República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(11) BR 112016030547-7 B1

(22) Data do Depósito: 06/11/2014

(45) Data de Concessão: 16/11/2022

(54) Título: SISTEMA E MÉTODO DE REPLICAÇÃO E DE ARMAZENAMENTO DISTRIBUÍDO

(51) Int.Cl.: G06F 9/50.

(73) Titular(es): HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD..

(72) Inventor(es): DAOHUI WANG; FENG ZHANG; XUYOU LIU.

(86) Pedido PCT: PCT CN2014090445 de 06/11/2014

(87) Publicação PCT: WO 2016/070375 de 12/05/2016

(85) Data do Início da Fase Nacional: 26/12/2016

(57) Resumo: SISTEMA E MÉTODO DE REPLICAÇÃO E DE ARMAZENAMENTO DISTRIBUÍDO. A presente invenção revela um sistema e um método de replicação e armazenamento distribuídos. O sistema inclui pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204) e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para configurar, para cada nó de OSD (206), pelo menos uma partição lógica (partição) correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206); o módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados correspondentes à solicitação de IO, em que o MDC (202) determina que um nó de OSD (206) no sistema é um nó de OSD (206) defeituoso, determina uma partição no nó de OSD (206) defeituoso, atualiza uma visualização de partição de um grupo de partições que inclui a partição no nó de OSD (206) defeituoso e envia uma notificação de atualização para um nó de OSD (206) primário na visualização de partição atualizada; e o nó de OSD (206) primário é (...).

“SISTEMA E MÉTODO DE REPLICAÇÃO E DE ARMAZENAMENTO DISTRIBUÍDO"

CAMPO DA TÉCNICA

[0001] A presente invenção refere-se ao campo de tecnologia da informação IT (Tecnologia da Informação) e, em particular, a um armazenamento distribuído e a um sistema e um método de replicação.

ANTECEDENTES

[0002] Com o rápido desenvolvimento de tecnologias da informação e a ampla aplicação da Internet, os dados gerados pelas pessoas aumentam de uma forma explosiva, o que impõe uma exigência maior sobre a capacidade de extensão de armazenamento de dados. Em comparação com um sistema de matriz de armazenamento convencional, um sistema de armazenamento distribuído tem uma capacidade de extensão melhor e uma compatibilidade com dispositivo de hardware comum, e pode corresponder melhor a uma exigência para armazenamento de dados posteriormente.

[0003] No sistema de armazenamento distribuído, geralmente, uma grande quantidade de nós de armazenamento é organizada para formar um sistema distribuído e a confiabilidade de dados é assegurada por meio de replicação e cópia de segurança de dados entre nós diferentes, de forma que os dados tenham réplicas em todos os nós de armazenamento diferentes. A forma como assegurar a consistência de dados de múltiplas réplicas de dados se tornou um problema que confronta o sistema de armazenamento distribuído por muito tempo. Em um caso de assegurar a consistência de dados, o desempenho e a disponibilidade do sistema também se tornam considerações de importância crescente.

[0004] A Figura 1 mostra o protocolo de comprometimento de duas fases existente (2 Fase Commit, 2PC), que é um protocolo de controle de réplica de consistência forte centralizado típico e é usado em muitos sistemas de banco de dados distribuídos para assegurar consistência de réplica.

[0005] No protocolo de comprometimento de duas fases, um sistema inclui geralmente dois tipos de nós: um coordenador (coordenador) e um participante (participante). O coordenador é responsável pela execução do início da votação a respeito da atualização de dados e notificar uma decisão de voto, e o participante participa na votação a respeito da atualização de dados e

executa a decisão de voto.

[0006] O protocolo de comprometimento de duas fases inclui duas fases: a Fase 1 é uma fase de solicitação de comprometimento, em que o coordenador instrui o participante a votar na modificação de dados e o participante notifica o coordenador de seu próprio resultado de votação: Sim ou Não; e a Fase 2 é uma fase de comprometimento, em que o coordenador faz uma decisão: comprometer ou Abortar, de acordo com o resultado de votação na primeira fase.

[0007] Executar de modo bem-sucedido o protocolo de comprometimento de duas fases, uma vez que exige pelo menos duas rodadas de interação entre o coordenador e cada participante com quatro mensagens e tempos excessivos de interação, degrada o desempenho. Além disso, no protocolo de comprometimento de duas fases, se um nó se tornar defeituoso ou não tiver resposta continuamente, outra solicitação de IO é bloqueada e finalmente falha devido a um tempo fora e uma reversão de dados precisa ser realizada. O protocolo de comprometimento de duas fases tem tolerância e disponibilidade de defeito relativamente baixas.

SUMÁRIO

[0008] As modalidades da presente invenção fornecem um sistema e um método de replicação e de armazenamento distribuído para gerenciar o armazenamento e a replicação de dados em um sistema de armazenamento distribuído, a fim de resolver um problema de baixo desempenho e baixa disponibilidade de um protocolo de replicação de consistência existente.

[0009] De acordo com um primeiro aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciada por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição

secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, armazenamento de dados que correspondem à solicitação de IO, em que o MDC determina que um nó de OSD no sistema é um nó de OSD defeituoso, determina uma partição no nó de OSD defeituoso, atualiza uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso e envia uma notificação de atualização para um nó de OSD primário na visualização de partição atualizada; e o nó de OSD primário é adaptado para processar, de acordo com a visualização de partição atualizada, após receber a notificação de atualização enviada pelo módulo de MDC, a replicação dos dados que correspondem à solicitação de IO.

[0010] Em uma primeira forma de implantação possível do primeiro aspecto, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD primário; e o processamento, de acordo com a visualização de partição atualizada, da replicação dos dados que correspondem à solicitação de IO inclui especificamente: replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados que correspondem à solicitação de IO a partir do módulo de roteamento de IO em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada ou em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD secundário cuja situação de partição é

inconsistente na visualização de partição localmente armazenada atualizada, mas que está recuperando dados.

[0011] Com referência à primeira forma de implantação possível do primeiro aspecto, em uma segunda forma de implantação possível, o módulo de MDC é adicionalmente adaptado para: gerar uma visualização de IO, em que a visualização de IO inclui um identificador de um nó de OSD primário de um grupo de partição; e enviar a visualização de IO para o módulo de roteamento de IO e o nó de OSD em que a partição na visualização de partição está localizada; e o nó de OSD primário é adicionalmente adaptado para: atualizar, de acordo com a visualização de partição atualizada, uma visualização de IO localmente armazenada no nó de OSD primário, e processar, de acordo com a visualização de IO localmente armazenada atualizada, uma replicação dos dados que correspondem à solicitação de IO.

[0012] Com referência à segunda forma de implantação possível do primeiro aspecto, em uma terceira forma de implantação possível, o módulo de MDC é adicionalmente adaptado para: quando se determina que a partição no nó de OSD defeituoso inclui uma partição primária, atualizar uma visualização de IO de um grupo de partição que inclui a partição primária e notificar um nó de OSD secundário na visualização de partição atualizada da visualização de IO atualizada; e o nó de OSD secundário na visualização de partição atualizada é adaptado para: atualizar uma visualização de IO localmente armazenada de acordo com a visualização de IO atualizada e processar, de acordo com a visualização de IO localmente armazenada atualizada, a replicação dos dados que correspondem à solicitação de IO.

[0013] Com referência à terceira forma de implantação possível do primeiro aspecto, em uma quarta forma de implantação possível, a atualização de uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso inclui especificamente: quando a partição no nó de OSD defeituoso inclui uma partição secundária, marcar uma situação de partição do nó de OSD defeituoso em uma visualização de partição de um grupo de partição que inclui a partição secundária como inconsistente; e, quando a partição no nó de OSD defeituoso inclui a partição primária, definir o nó de OSD defeituoso que serve como um nó de OSD primário em uma visualização de partição do grupo de partição que inclui a partição primária como um novo nó de

OSD secundário, marcar uma situação de partição que corresponde ao novo nó de OSD secundário como inconsistente, selecionar um nó de OSD secundário cuja situação de partição é consistente a partir de um nó de OSD secundário original na visualização de partição do grupo de partição que inclui a partição primária e definir o nó de OSD secundário selecionado como um novo nó de OSD primário.

[0014] De acordo com um segundo aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar armazenamento de dados de IO de acordo com a solicitação de IO, em que o módulo de roteamento de IO é adaptado para: receber a solicitação de IO, em que a solicitação de IO inclui uma chave; determinar, de acordo com a chave, um grupo de partição ao qual os dados que correspondem à solicitação de IO pertencem, e determinar um nó de OSD primário do grupo de partição ao qual os dados pertencem; adicionar informações de versão de visualização de IO de uma visualização de

IO do grupo de partição ao qual os dados pertencem à solicitação de IO; e enviar, para o nó de OSD primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO; o nó de OSD primário é adaptado para receber a solicitação de IO; executar a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD primário; gerar uma solicitação de replicação que porta as informações de versão de visualização de IO; e enviar a solicitação de replicação para um nó de OSD secundário do grupo de partição ao qual os dados pertencem; e o nó de OSD secundário é adaptado para: receber a solicitação de replicação e executar a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD secundário, de forma que os dados que correspondem à solicitação de IO no nó de OSD secundário se mantenham consistentes com os dados que correspondem à solicitação de IO no nó de OSD primário.

[0015] Em uma primeira forma de implantação possível do segundo aspecto, o nó de OSD primário é adicionalmente adaptado para: retornar um erro para o módulo de roteamento de IO após determinar, de acordo com as informações de versão de visualização de IO, que a versão de visualização de IO na solicitação de IO é anterior à versão de visualização de IO localmente armazenada no OSD primário; e, após determinar que a versão de visualização de IO na solicitação de IO é posterior à versão de visualização de IO localmente armazenada no OSD primário, adicionar a solicitação de IO a uma fila de cache e consultar o módulo de MDC sobre as informações de versão de visualização de IO das visualização de IO do grupo de partição ao qual os dados pertencem, a fim de executar a solicitação de IO após determinar que a versão de visualização de IO localmente armazenada no OSD primário é consistente com a versão de visualização de IO na solicitação de IO; e o módulo de roteamento de IO é adaptado para: após receber o erro retornado pelo nó de OSD primário, consultar o módulo de MDC sobre a visualização de IO do grupo de partição ao qual os dados pertencem e, após obter informações de versão de visualização de IO atualizada, enviar uma solicitação de IO que porta as informações de

versão de visualização de IO atualizada.

[0016] Com referência ao segundo aspecto ou à primeira forma de implantação possível do segundo aspecto, em uma segunda forma de implantação possível, as informações de versão de visualização de IO incluem um número de versão de visualização de IO e o nó de OSD primário gera adicionalmente um identificador de sequência para a solicitação de IO e adiciona o identificador de sequência à solicitação de replicação enviada para o nó de OSD secundário, em que o identificador de sequência inclui o número de versão de visualização de IO e um número de sequência, e o número de sequência indica um número de série de uma operação de modificação de dados que correspondem a um grupo de partição na visualização de IO dentro de uma versão de visualização de IO; e o nó de OSD secundário é adicionalmente adaptado para executar a solicitação de replicação de acordo com o identificador de sequência.

[0017] Com referência à segunda forma de implantação possível do segundo aspecto, em uma terceira forma de implantação possível, a solicitação de replicação porta adicionalmente um identificador de sequência em uma solicitação de replicação anterior enviada pelo nó de OSD primário para o grupo de partição e o nó de OSD secundário é adaptado para: após receber a solicitação de replicação, executar a solicitação de replicação quando o identificador de sequência na solicitação de replicação anterior é consistente com um identificador de sequência maior localmente armazenado no nó de OSD secundário.

[0018] Com referência ao segundo aspecto ou à primeira, à segunda e à terceira formas de implantação possíveis do segundo aspecto, em uma quarta forma de implantação possível, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o módulo de MDC é adicionalmente adaptado para: quando se detecta, em um processo de processamento da solicitação de IO, que o nó de OSD primário se torna defeituoso, definir o nó de OSD primário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD secundário e marcar uma situação de partição do novo OSD secundário como inconsistente, definir qualquer nó de OSD secundário do nó de

OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD primário, notificar o novo nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e atualizar, usando-se o novo nó de OSD primário, a visualização de IO do grupo de partição ao qual os dados pertencem e notificar o módulo de roteamento de IO da visualização de IO atualizada da partição à qual os dados pertencem; o módulo de roteamento de IO é adicionalmente adaptado para: receber a visualização de IO atualizada que é do grupo de partição e que é enviada pelo módulo de MDC e enviar a solicitação de IO para o novo nó de OSD primário de acordo com a visualização de IO atualizada do grupo de partição; e o novo nó de OSD primário é adaptado para: receber a solicitação de IO e, após executar a solicitação de IO, gerar uma segunda solicitação de replicação e enviar a segunda solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada do grupo de partição ao qual os dados pertencem.

[0019] Com referência ao segundo aspecto ou à primeira, à segunda e à terceira formas de implantação possíveis do segundo aspecto, em uma quinta forma de implantação possível, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o módulo de MDC é adicionalmente adaptado para: quando se detecta, em um processo de processamento da solicitação de IO, que qualquer nó de OSD secundário do nó de OSD secundário se torna defeituoso, marcar uma situação de partição do qualquer OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como inconsistente e notificar o nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e o nó de OSD primário é adaptado para: após receber a visualização de partição atualizada do grupo de partição ao qual os dados pertencem, enviar a solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada e omitir o envio da solicitação de replicação para o nó de OSD secundário cuja situação de partição é inconsistente.

[0020] De acordo com um terceiro aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo

menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO, e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, armazenamento de dados que correspondem à solicitação de IO, em que o nó de OSD é adaptado para: enviar uma solicitação de consulta para o módulo de MDC após um failback para solicitar uma visualização de IO de um grupo de partição que inclui uma partição no nó de OSD, em que o nó de OSD é chamado um nó de OSD de failback e a solicitação de consulta porta um identificador de OSD do nó de OSD de failback; receber a visualização de IO retornada pelo MDC; iniciar uma solicitação de recuperação de dados para um OSD primário na visualização de IO para solicitar a recuperação de dados atualizados pelo nó de OSD de failback durante uma falha; receber os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processar a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados; o módulo de MDC é adaptado para: receber a solicitação de consulta do nó de OSD de

failback, retornar a visualização de IO para o nó de OSD de failback de acordo com o identificador de OSD na solicitação de consulta e atualizar a visualização de partição do grupo de partição após o nó de OSD de failback concluir a recuperação de dados; e o nó de OSD primário é adaptado para: receber a solicitação de recuperação de dados do nó de OSD de failback, enviar os dados atualizados durante a falha para o nó de OSD de failback e processar, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados, a replicação de dados que correspondem à solicitação de IO.

[0021] Em uma primeira forma de implantação possível do terceiro aspecto, o nó de OSD primário é adicionalmente adaptado para: após receber a solicitação de recuperação de dados, receber a solicitação de IO enviada pelo módulo de roteamento de IO para a partição no nó de OSD de failback, executar a solicitação de IO e enviar, para o nó de OSD de failback, uma solicitação de replicação que porta informações de chave de IO e os dados que correspondem à solicitação de IO; e o nó de OSD de failback escreve, para um registro, a solicitação de replicação que porta as informações de chave de IO e os dados que correspondem à solicitação de IO e atualizar, de acordo com uma gravação do registro, após a recuperação de dados ser concluída, os dados que correspondem à solicitação de IO para recursos de armazenamento físicos gerenciados pelo nó de OSD de failback.

[0022] Com referência à primeira forma de implantação possível do terceiro aspecto, em uma segunda forma de implantação possível, a solicitação de recuperação de dados porta um identificador de sequência maior que é de uma operação de IO para a partição no nó de OSD de failback e que é localmente gravada no nó de OSD de failback, em que o identificador de sequência maior é um último número de versão de visualização de IO da visualização de IO do grupo de partição que inclui a partição no nó de OSD de failback e um número de série maior de uma operação de modificação de dados que correspondem a uma partição na visualização de IO que corresponde ao último número de versão de visualização de IO; e o envio dos dados atualizados durante a falha ao nó de OSD de failback inclui: determinar que o identificador de sequência maior na solicitação de recuperação de dados é maior ou igual a um identificador de sequência menor atual localmente armazenado no nó de OSD primário, enviar

uma entrada em que o nó de OSD de failback não está presente durante a falha ao nó de OSD de failback, receber uma solicitação de recuperação de dados iniciada pelo nó de OSD de failback de acordo com a entrada e enviar dados que correspondem à entrada para o nó de OSD de failback, em que o menor identificador de sequência é um número de versão de visualização de IO menor que é da visualização de IO do grupo de partição e que é armazenado no nó de OSD primário e um número de série menor de uma operação de modificação em dados que correspondem a uma partição na visualização de IO que corresponde ao menor número de versão de visualização de IO.

[0023] Com referência à primeira forma de implantação possível do terceiro aspecto, em uma terceira forma de implantação possível, a solicitação de recuperação de dados porta um identificador de sequência maior que é de uma operação de IO para partição no nó de OSD de failback e que é localmente gravado no nó de OSD de failback, em que o identificador de sequência maior inclui: um último número de versão de visualização de IO da visualização de IO do grupo de partição que inclui a partição no nó de OSD de failback e um número de série maior de uma operação de modificação em dados que correspondem a uma partição na visualização de IO dentro da visualização de IO que corresponde ao último número de versão de visualização de IO; e o envio dos dados atualizados durante a falha para o nó de OSD de failback inclui: determinar que o identificador de sequência maior na solicitação de recuperação de dados é menor que um identificador de sequência menor atual localmente armazenado no nó de OSD primário, enviar o identificador de sequência menor atual localmente armazenado no OSD primário para o nó de OSD de failback, receber uma solicitação de recuperação de dados, iniciada pelo nó de OSD de failback, para sincronizar todos os dados que correspondem a uma partição primária que pertence ao grupo de partição e que está no nó de OSD primário e enviar todos os dados que correspondem à partição primária para o nó de OSD de failback, em que o menor identificador de sequência é um número de versão de visualização de IO menor que é da visualização de IO do grupo de partição e que é armazenado no nó de OSD primário e um número de série menor de uma operação de modificação em dados que correspondem a uma partição na visualização de IO que corresponde ao menor número de versão de visualização de IO.

[0024] De acordo com um quarto aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde aos recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD, em que uma partição em um grupo de partição está localizada; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados que correspondem à solicitação de IO, em que o sistema inclui uma memória e um processador, em que a memória é adaptada para armazenar uma instrução legível por computador e a instrução é usada para executar funções do módulo de MDC, do módulo de roteamento de IO e do nó de OSD; e o processador é adaptado para: ser conectado à memória, ler a instrução na memória, e permitir, de acordo com a instrução, que o processador execute as seguintes operações: determinar que um nó de OSD no sistema é um nó de OSD defeituoso, determinar uma partição no nó de OSD defeituoso, atualizar uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso e enviar uma notificação de atualização para um nó de OSD primário em que o grupo de partição na visualização de partição atualizada está localizado, de forma que o nó de OSD primário processe, de acordo com a visualização de partição atualizada, a replicação de dados que correspondem à

solicitação de IO.

[0025] De acordo com um quinto aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO, e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com uma solicitação de IO, o armazenamento de dados que correspondem à solicitação de IO, em que o sistema inclui uma memória e um processador, em que uma memória é adaptada para armazenar uma instrução legível por computador e uma instrução é usada para executar funções do módulo de MDC, do módulo de roteamento de IO e do nó de OSD; e o processador é adaptado para: ser conectado à memória, ler a instrução na memória e permitir, de acordo com a instrução, que o processador execute as seguintes operações: permitir que o módulo de roteamento de IO: receba a solicitação de IO, em que a solicitação de IO inclui uma chave; determine, de acordo com a chave, um grupo de partição ao qual dados que correspondem à solicitação de IO pertencem e determine um nó de OSD primário do grupo de partição ao qual os dados pertencem; adicionar

informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem à solicitação de IO; e enviar, para o nó de OSD primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO; permitir que o nó de OSD primário: receba a solicitação de IO; execute a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO localmente armazenada; gere uma solicitação de replicação que porta as informações de versão de visualização de IO; e envie a solicitação de replicação para um nó de OSD secundário da partição à qual os dados pertencem; e permite que o nó de OSD secundário: receba a solicitação de replicação e execute a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD secundário, de forma que dados que correspondem à solicitação de IO no nó de OSD secundário se mantenha consistente com dados que correspondem à solicitação de IO no nó de OSD primário.

[0026] De acordo com um sexto aspecto, um sistema de armazenamento e replicação distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, em que uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar

uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados que correspondem à solicitação de IO, em que o sistema inclui uma memória e um processador, em que a memória é adaptada para armazenar uma instrução legível por computador e a instrução é usada para executar funções do módulo de MDC, do módulo de roteamento de IO e do nó de OSD; e o processador é adaptado para: ser conectado à memória, ler a instrução na memória e permitir, de acordo com a instrução, que o processador execute as seguintes operações: permitir que o nó de OSD: envie uma solicitação de consulta para o módulo de MDC após um failback para solicitar uma visualização de IO de um grupo de partição que inclui uma partição no nó de OSD, em que o nó de OSD é chamado um nó de OSD de failback e a solicitação de consulta porta um identificador de OSD do nó de OSD de failback; receba a visualização de IO retornada pelo MDC; inicie uma solicitação de recuperação de dados para um OSD primário na visualização de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD de failback durante uma falha; receba os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processe a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados; permitir que o módulo de MDC: receba a solicitação de consulta do nó de OSD de failback, retorne a visualização de IO para o nó de OSD de failback de acordo com o identificador de OSD na solicitação de consulta e atualize a visualização de partição do grupo de partição após o nó de OSD de failback concluir a recuperação de dados; e permitir que o nó de OSD primário: receba a solicitação de recuperação de dados do nó de OSD de failback, envie os dados atualizados durante a falha para o nó de OSD de failback e processe, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados, a replicação dos dados que correspondem à solicitação

de IO.

[0027] De acordo com um sétimo aspecto, um método para gerenciar armazenamento e replicação de dados em um sistema de armazenamento distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físicos gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primária e secundária, uma partição primária e uma partição secundária que correspondem à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em nós de OSD diferentes, um nó de OSD em que uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária e um nó de OSD em que uma partição secundária está localizada um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição inclui informações sobre um OSD em que uma partição em um grupo de partição está localizada; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados que correspondem à solicitação de IO, em que o método inclui: determinar que um nó de OSD no sistema é um nó de OSD defeituoso, determinar uma partição no nó de OSD defeituoso, atualizar uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso e enviar uma notificação de atualização para um nó de OSD primário na visualização de partição atualizada; e o nó de OSD primário é adaptado para processar, de acordo com a visualização de partição atualizada após receber a notificação de atualização enviada pelo módulo de MDC, a replicação dos dados que correspondem à solicitação de IO.

[0028] Em uma primeira forma de implantação possível do sétimo aspecto, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um

OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD primário; e o processamento, de acordo com a visualização de partição atualizada, da replicação dos dados que correspondem à solicitação de IO inclui especificamente: replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados que correspondem à solicitação de IO a partir do módulo de roteamento de IO em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada ou em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que está recuperando dados.

[0029] Com referência à primeira forma de implantação possível do sétimo aspecto, em uma segunda forma de implantação possível, o método inclui adicionalmente: quando se determina que a partição no nó de OSD defeituoso inclui uma partição primária, atualizar, pelo módulo de MDC, uma visualização de IO de um grupo de partição que inclui a partição primária e notificar um nó de OSD secundário na visualização de partição atualizada da visualização de IO atualizada; e atualizar, pelo nó de OSD secundário na visualização de partição atualizada, uma visualização de IO localmente armazenada de acordo com a visualização de IO atualizada e processar, de acordo com a visualização de IO localmente armazenada atualizada, a replicação dos dados que correspondem à solicitação de IO.

[0030] Com referência à segunda forma de implantação possível do sétimo aspecto, em uma terceira forma de implantação possível, a atualização de uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso inclui especificamente: quando a partição no nó de OSD defeituoso inclui uma partição secundária, marcar uma situação de partição do nó de OSD defeituoso em uma visualização de partição de um grupo de partição que inclui a partição secundária como inconsistente; e, quando a partição no nó de OSD defeituoso inclui a partição primária, definir o nó de OSD defeituoso que serve como um nó de OSD primário em uma visualização de

partição do grupo de partição que inclui a partição primária como um novo nó de OSD secundário, marcar uma situação de partição que corresponde ao novo nó de OSD secundário como inconsistente, selecionar um nó de OSD secundário cuja situação de partição é consistente a partir de um nó de OSD secundário original na visualização de partição do grupo de partição que inclui a partição primária e definir o nó de OSD secundário selecionado como um novo nó de OSD primário.

[0031] De acordo com um oitavo aspecto, um método para gerenciar armazenamento de dados e replicação em um sistema de armazenamento distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para: configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físico gerenciados por cada nó de OSD, em que a pelo menos uma partição é uma partição primária, uma partição secundária ou qualquer combinação de partições primárias e secundárias, uma partição primária e uma partição secundária que corresponde à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD, um nó de OSD no qual uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária, e um nó de OSD no qual uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição inclui informações a respeito de um OSD no qual uma partição em um grupo de partição está localizada, e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar os armazenamento de dados de IO de acordo com a solicitação de IO, em que o método inclui: o módulo de roteamento de IO é adaptado para: receber a solicitação de IO, em que a solicitação de IO inclui uma chave; determinar, de acordo com a chave, um grupo de partição ao qual os dados que correspondem à solicitação de IO pertencem, e determinar um nó

de OSD primário do grupo de partição ao qual os dados pertencem; adicionar informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem para a solicitação de IO; e enviar, para o nó de OSD primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO; o nó de OSD primário é adaptado para: receber a solicitação de IO; executar a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO armazenada localmente no nó de OSD primário; gerar uma solicitação de replicação que porta as informações de versão de visualização de IO; e enviar a solicitação de replicação para um nó de OSD secundário do grupo de partição ao qual os dados pertencem; e o nó de OSD secundário é adaptado para: receber a solicitação de replicação e executar a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO armazenada localmente no nó de OSD secundário, de modo que os dados que correspondem à solicitação de IO no nó de OSD secundário se mantêm consistentes em relação aos dados que correspondem à solicitação de IO no nó de OSD primário.

[0032] Com referência a um primeiro modo de implantação possível do oitavo aspecto, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD no qual uma partição em um grupo de partição está localizada; e o método inclui adicionalmente: durante a detecção, em um processo de processamento da solicitação de IO, que o nó de OSD primário se tornou defeituoso, definir, através do módulo de MDC, o nó de OSD primário em uma visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD secundário, marcando uma situação de partição do novo OSD secundário como inconsistente; definir qualquer nó de OSD secundário do nó de OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD primário, notificar o novo nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e atualizar, através da utilização do novo nó de OSD primário, a visualização de IO do grupo de partição ao qual os dados pertencem,

e notificar o módulo de roteamento de IO da visualização de IO atualizada da partição à qual os dados pertencem; o módulo de roteamento de IO é adicionalmente adaptado para: receber a visualização de IO atualizada que é do grupo de partição e que é enviada pelo módulo de MDC, e enviar a solicitação de IO ao novo nó de OSD primário de acordo com a visualização de IO atualizada do grupo de partição; e o novo nó de OSD primário é adaptado para: receber a solicitação de IO e, após executar a solicitação de IO, gerar uma segunda solicitação de replicação, e enviar a segunda solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada do grupo de partição ao qual os dados pertencem.

[0033] Com referência a um segundo modo de implantação possível do segundo aspecto, a visualização de partição inclui especificamente uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD no qual uma partição em um grupo de partição está localizada; e o método inclui adicionalmente: durante a detecção, em um processo de processamento da solicitação de IO, que qualquer nó de OSD secundário do nó de OSD secundário se tornou defeituoso, marcar, através do módulo de MDC, uma situação de partição de qualquer nó de OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como inconsistente, e notificar o nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e o nó de OSD primário é adaptado para: após receber a visualização de partição atualizada do grupo de partição ao qual os dados pertencem, enviar a solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada, e ignorar o envio da solicitação de replicação para o nó de OSD secundário cuja situação de partição é inconsistente.

[0034] De acordo com um novo aspecto, um método para gerenciar o armazenamento de dados e a replicação em um sistema de armazenamento distribuído é fornecido, em que o sistema inclui pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para: configurar, para cada nó de OSD, pelo menos uma partição lógica (partition) que corresponde a recursos de armazenamento físico gerenciados por cada nó de OSD, em que pelo menos

uma partição é uma partição primária, uma partição secundária, ou qualquer combinação de partições primárias e secundárias, uma partição primária e uma partição secundária que corresponde à partição primária constituem um grupo de partição, uma partição primária e uma partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD, um nó de OSD no qual uma partição primária está localizada é um nó de OSD primário de um grupo de partição que inclui a partição primária, e um nó de OSD no qual uma partição secundária está localizada é um nó de OSD secundário de um grupo de partição que inclui a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com uma partição, em que a visualização de partição inclui informações a respeito de um OSD no qual uma partição em um grupo de partição está localizada, e a visualização de IO inclui um identificador de um OSD primário de um grupo de partição; o módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados que correspondem à solicitação de IO, em que o método inclui: o nó de OSD ser adaptado para: enviar uma solicitação de consulta para o módulo de MDC após um failback para solicitar um failback para solicitar uma visualização de IO de um grupo de partição que inclui uma partição no nó de OSD, em que o nó de OSD é chamado de um nó de OSD de failback, e a solicitação de consulta porta um identificador de OSD do nó de OSD de failback; receber a visualização de IO retornada pelo MDC; iniciar a solicitação de recuperação de dados para um OSD primário na visualização de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD de failback durante uma falha; receber os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processar a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados; o módulo de MDC é adaptado para: receber a solicitação de consulta do nó de OSD de failback, retornar a visualização de IO ao nó de OSD de failback de acordo com o identificador de OSD na solicitação de consulta, e atualizar a visualização de partição do grupo de partição após o nó de OSD de failback concluir a recuperação de dados; e o nó de OSD primário é adaptado para: receber uma solicitação de recuperação de dados do nó de OSD

de failback, enviar os dados atualizados durante a falha para o nó de OSD de failback, e processar, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback concluir a recuperação de dados, a replicação de dados que correspondem à solicitação de IO.

BREVE DESCRIÇÃO DOS DESENHOS

[0035] Para descrever as soluções técnicas nas modalidades da presente invenção mais claramente, a seguir introduz-se brevemente os desenhos anexos exigidos para descrever as modalidades. Aparentemente, os desenhos anexos na descrição a seguir mostram meramente algumas modalidades da presente invenção e um indivíduo de habilidade comum na técnica ainda pode derivar outros desenhos a partir desses desenhos anexos sem esforços criativos.

[0036] A Figura 1 é um fluxograma do protocolo de comprometimento de duas fases da técnica anterior;

[0037] A Figura 2A é um diagrama de arquitetura esquemática de um sistema de armazenamento e replicação distribuído de acordo com uma modalidade da presente invenção;

[0038] A Figura 2B é um diagrama estrutural esquemático de um sistema de armazenamento e replicação distribuído de acordo com outra modalidade da presente invenção;

[0039] A Figura 2C é um diagrama estrutural esquemático de um sistema de armazenamento e replicação distribuído de acordo com outra modalidade da presente invenção;

[0040] A Figura 3 é um diagrama esquemático de uma visualização de agrupamento de acordo com uma modalidade da presente invenção;

[0041] A Figura 4 é um diagrama esquemático de uma transição de situação de visualização de OSD de acordo com uma modalidade da presente invenção;

[0042] A Figura 5 é um diagrama estrutural esquemático de um sistema de armazenamento e replicação distribuído de acordo com outra modalidade da presente invenção;

[0043] A Figura 6 é um fluxograma de inicialização de visualização de acordo com uma modalidade da presente invenção;

[0044] A Figura 7A e a Figura 7B são um fluxograma de processamento de solicitação de IO de acordo com uma modalidade da presente invenção;

[0045] A Figura 8A e a Figura 8B são um fluxograma de processamento de falha de OSD de acordo com uma modalidade da presente invenção;

[0046] A Figura 9A, a Figura 9B e a Figura 9C são um fluxograma de processamento de failback de OSD de acordo com uma modalidade da presente invenção;

[0047] A Figura 10A e a Figura 10B são um fluxograma de recuperação de dados em um processo de processamento de failback de OSD de acordo com uma modalidade da presente invenção;

[0048] A Figura 11A e a Figura 11B são um fluxograma de processamento realizado após um OSD sair de um agrupamento de acordo com uma modalidade da presente invenção; e

[0049] A Figura 12A e a Figura 12B são um fluxograma de processamento realizado após um novo OSD se unir a um agrupamento de acordo com uma modalidade da presente invenção.

DESCRIÇÃO DAS MODALIDADES

[0050] A seguir descreve-se clara e completamente as soluções técnicas nas modalidades da presente invenção com referência aos desenhos anexos nas modalidades da presente invenção. Aparentemente, as modalidades descritas são meramente algumas, mas não todas as modalidades da presente invenção. Todas as outras modalidades obtidas por um indivíduo de habilidade comum na técnica com base nas modalidades da presente invenção sem esforços criativos devem ser abrangidas pelo escopo de proteção da presente invenção.

[0051] Conforme mostrado na Figura 2A, uma modalidade específica da presente invenção fornece um sistema de controle de armazenamento e replicação distribuído, de modo a gerenciar e controlar o armazenamento de dados e a replicação mencionados nessa modalidade da presente invenção. O sistema de controle de armazenamento e replicação distribuído inclui principalmente três subcamadas: uma camada de situação, uma camada de interface e uma camada de dados. A camada de situação inclui

um módulo de Controlador de metadados (MDC) de controle de metadados 202, e em uma aplicação real, o fato de se um MDC secundário de um MDC precisa ser configurado e uma quantidade de MDCs secundários podem ser determinados de acordo com uma exigência, em que o MDC secundário é adaptado para fazer o papel de um MDC primário quando o módulo de MDC primário se torna defeituoso. A camada de interface inclui múltiplos módulos de roteamento de IO (módulo de roteamento de entrada/saída) 204 (que também podem ser chamados de clientes, e os dois conceitos podem ser intercambiáveis na implantação da presente invenção). A camada de dados inclui múltiplos nós de dispositivo de armazenamento de objeto (OSD) de dispositivo de armazenamento baseado em objeto 206. A camada de situação se comunica com a camada de interface e a camada de dados através do uso de uma mensagem de visualização de situação. Por exemplo, o módulo de MDC 202 envia uma notificação de atualização ao módulo de roteamento de IO 204 e ao nó de OSD 206 através do uso da mensagem de visualização de situação, de modo a instruir o módulo de roteamento de IO 204 e o nó de OSD 206 para atualizar uma visualização de agrupamento local (que também pode ser chamada de uma visualização, e os dois conceitos podem ser intercambiáveis na implantação da presente invenção); ou envia diretamente uma visualização de agrupamento gerada ou atualizada pelo módulo de MDC 202 para o módulo de roteamento de IO 204 e o nó de OSD 206. A camada de interface e a camada de dados se comunicam uma com a outra através do uso de uma mensagem de serviço. Por exemplo, o módulo de roteamento de IO 204 envia uma mensagem de solicitação de IO para o nó de OSD 206 para solicitar o armazenamento de dados de IO e a replicação.

[0052] O módulo de MDC 202, como uma entrada para a entrega de informações de configuração de agrupamento, é adaptado para alocar uma partição lógica (partition) de recursos de armazenamento lógico no espaço de armazenamento de aplicativo para cada nó de OSD, gerar uma visualização de agrupamento de acordo com a partição, manter e atualizar a visualização de agrupamento, e notificar o módulo de roteamento de IO correspondente 204 e o nó de OSD 206 da atualização de visualização de agrupamento.

[0053] O módulo de roteamento de IO 204 é adaptado para rotear e encaminhar uma solicitação de IO de um aplicativo de camada superior para

um nó de ODS correspondente de acordo com a visualização de agrupamento.

[0054] O nó de OSD 206 é adaptado para: executar uma operação de IO relacionada na solicitação de IO de acordo com a visualização de agrupamento, em que a operação de IO relacionada inclui principalmente dados de armazenamento e replicação, para implantar uma consistência de cópia de segurança de dados; e organizar uma operação de dados em recursos de armazenamento físico (por exemplo, disco local ou recursos de armazenamento externo) gerenciados pelo nó de OSD 206.

[0055] Pode-se compreender que o módulo de MDC supracitado, o módulo de roteamento de IO e o nó de OSD podem ser implantados por hardware, firmware, software ou uma combinação dos mesmos. Em uma aplicação real, um modo de implantação específico é determinado em consideração de uma exigência de projeto ou custos de fabricação de produto, e a presente invenção não se limita a um modo de implantação específico.

[0056] Em uma modalidade específica da presente invenção, todo o sistema de armazenamento e replicação distribuído pode ser empregado em uma plataforma ou servidor independente (por exemplo, uma plataforma de armazenamento e replicação distribuída na Figura 2B antecedente), de modo a se gerenciar o armazenamento e replicação de dados em um sistema de armazenamento distribuído conectado à plataforma ou servidor.

[0057] Em outra modalidade específica da presente invenção, o sistema de controle de armazenamento e replicação distribuído pode ser empregado, de um modo distribuído, em um sistema de armazenamento distribuído mostrado na Figura 2C. O sistema de armazenamento distribuído inclui múltiplos servidores ou hospedeiros, em que o hospedeiro ou servidor nessa modalidade é um hospedeiro ou servidor físico, ou seja, inclui hardware, tal como um processador e uma memória. O módulo de MDC 202 anterior pode ser empregado apenas em um servidor ou hospedeiro (sem MDC secundário), ou em dois servidores ou hospedeiros (um módulo de MDC primário e um módulo de MDC secundário), ou em três servidores ou hospedeiros (um módulo de MDC primário e dois módulos de MDC secundários) no sistema de armazenamento distribuído; o módulo de roteamento de IO 204 é empregado em cada servidor ou hospedeiro no sistema de armazenamento distribuído; e o nó de OSD 206 é empregado em cada servidor ou hospedeiro que tem recursos de

armazenamento no sistema de armazenamento distribuído, de modo a gerenciar e controlar recursos de armazenamento locais ou recursos de armazenamento externo. Em uma aplicação real, ou o módulo de roteamento de IO ou o nó de OSD, ou ambos dentre o módulo de roteamento de IO e o nó de OSD podem ser empregados em um hospedeiro, em um modo de emprego específico pode ser determinado de acordo com uma situação específica real, que não é limitada na presente invenção. O módulo de MDC 202, o módulo de roteamento de IO 204 e o nó de OSD 206 na Figura 2C constituem um sistema de controle de armazenamento distribuído, que é chamado de uma camada de protocolo de replicação distribuída no sistema de armazenamento distribuído mostrado na Figura 2B. O sistema de armazenamento distribuído controla o armazenamento e a replicação de dados de IO para recursos de armazenamento na camada de armazenamento através do uso da camada de protocolo de replicação distribuída. A camada de armazenamento inclui recursos de armazenamento local nos múltiplos servidores ou hospedeiros, e os módulos que se encontram na camada de protocolo de replicação distribuída e são distribuídos no servidor ou hospedeiro interagem uns com os outros através do uso de uma rede de dados comutados na camada de rede. Em um modo de implantação específico, a Ethernet ou infiniband podem ser usadas. Deve-se compreender que a Ethernet ou infiniband anterior é meramente um modo de implantação exemplificador de uma rede de dados comutados de alta velocidade usada nessa modalidade da presente invenção, que não é limitada nessa modalidade da presente invenção.

[0058] A seguir são usadas modalidades e modos de implantação específicos para se descrever a conexão e interação, funções específicas e semelhantes do módulo de MDC 202, do módulo de roteamento de IO 204 e do nó de OSD 206 no sistema de controle de armazenamento e replicação distribuído anterior em detalhes.

[0059] Em uma modalidade específica da presente invenção, uma função de particionamento do módulo de MDC pode incluir especificamente: o módulo de MDC configura, para cada nó de OSD de acordo com uma situação de recursos de armazenamento físico gerenciados por cada nó de OSD, uma partição lógica (partition) que corresponde aos recursos de armazenamento físico gerenciados por cada nó de OSD. A partição inclui uma quantidade

específica de blocos de dados no espaço de armazenamento de aplicativo. Em comparação ao espaço de armazenamento físico na camada de armazenamento, o espaço de armazenamento de aplicativo na camada de aplicativo é uma quantidade específica de espaço de armazenamento lógico alocada pela camada de aplicativo para um usuário, e é um mapeamento lógico do espaço de armazenamento físico na camada de armazenamento. Ou seja, um conceito da partição no presente documento é diferente de um conceito de uma partição de espaço de armazenamento físico. Quando os dados estão sendo armazenados, o espaço de uma partição no espaço de armazenamento de aplicativo pode ser mapeado para uma ou mais partições no espaço de armazenamento físico. Uma granularidade específica da partição pode ser adquirida a partir de informações de configuração de agrupamento, ou pode ser determinada pelo módulo de MDC de acordo com uma regra específica ou determinada de outro modo, que não é limitado na presente invenção.

[0060] Em um modo de implantação específico, o módulo de MDC pode gerar uma visualização de agrupamento da partição de acordo com informações, tais como informações de configuração de tamanho de partição, uma situação de recurso de armazenamento local e informações de uma situação de recurso de armazenamento externo (por exemplo, LUN (Logical Unit Number, número de unidade lógica) de uma SAN (Storage Area Network, rede de área de armazenamento) normalmente acessada).

[0061] De modo geral, para se assegurar a confiabilidade e disponibilidade de dados, uma partição tem réplicas armazenadas em diferentes nós de OSD, e uma quantidade de réplica de partição pode ser configurada com o uso de um arquivo de configuração, ou pode ser determinada pelo MDC de acordo com um algoritmo específico. Há uma partição primária e uma partição secundária por classificação: Uma réplica é selecionada a partir de múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de uma partição primária; uma réplica exceto a réplica primária da partição é chamada de uma partição secundária. Uma partição primária e uma partição secundária que corresponde à partição primária constituem um grupo de partição, um nó de OSD no qual uma partição primária está localizada é chamado de um nó de OSD primário de um grupo de partição que inclui na partição primária, e um OSD primário descrito nessa modalidade se refere a um

OSD primário de um grupo de partição; um nó de OSD no qual uma partição secundária está localizada é chamado de um nó de OSD secundário de um grupo de partição que inclui a partição secundária, e um OSD secundário descrito nessa modalidade se refere a um OSD secundário para um grupo de partição.

[0062] Para facilitar a compreensão, é fornecida a seguir uma descrição adicional com referência à modalidade fornecida na Figura 2C. Conforme mostrado na Figura 2C, os recursos de armazenamento gerenciados por um OSD em um hospedeiro ou servidor servidor_1 (os conceitos do hospedeiro e do servidor nessa modalidade da presente invenção podem ser intercambiáveis) são divididos em uma partição 1, uma partição 2 e uma partição 3 (P1, P2 e P3 de modo conciso) e uma partição 4', uma partição 5' e uma partição 6' (P4', P5' e P6' de modo conciso), em que a P4', a P5' e a P6' são réplicas de uma partição 4, uma partição 5 e uma partição 6 (P4, P5 e P6 de modo conciso) em um nó de OSD em um servidor servidor_2, respectivamente. Há uma relação de mapeamento correspondente entre as partições no OSD no servidor_1 e os recursos de armazenamento físico na camada de armazenamento. Por exemplo, o espaço de uma partição no OSD é mapeado para um ou mais Blocos no espaço de armazenamento físico.

[0063] O OSD no hospedeiro ou servidor servidor_1 gerencia as partições primárias (a P1, a P2 e a P3) e as partições secundárias (a P4' a P5' e a P6') e o ODS é separadamente um nó de OSD primário de um grupo de partição que inclui a P1 e uma P1', um nó de OSD primário de um grupo de partição que inclui a P2 e uma P2' e um nó de OSD primário de um grupo de partição que inclui a P3 e uma P3'; além disso, o OSD é separadamente um nó de OSD secundário de uma partição que inclui a P4 e a P4', um nó de OSD secundário de uma partição que inclui a P5 e a P5', e um nó de OSD secundário de uma partição que inclui a P6 e a P6'. Pode-se compreender que, para diferentes grupos de partição, um mesmo nó de OSD pode servir como um nó de OSD primário e um nó de OSD secundário ao mesmo tempo.

[0064] A supracitada partição e as réplicas correspondentes podem ser definidas de acordo com os fatores a seguir, e em uma aplicação real específica, outro fator pode ser levado em consideração para se definir e planejar uma partição de disco.

[0065] Primeiro, segurança de dados: as réplicas de cada partição devem ser distribuídas a diferentes hospedeiros ou servidores o máximo possível. Um ponto principal de segurança de dados é que não seja permitido que múltiplas réplicas de uma partição sejam colocadas em um mesmo hospedeiro ou servidor. Em segundo lugar, balanço de dados: Uma quantidade de partições em cada OSD é mantida igual o máximo possível. Uma quantidade de partições primárias, uma quantidade de partições secundárias 1 e uma quantidade de partições secundárias 2 em cada OSD são mantidas iguais o máximo possível, de modo que os serviços processados em todos os OSDs sejam balanceados, e não surja nenhum ponto de acesso. Em terceiro lugar, dispersão de dados: as réplicas de uma partição em cada OSD devem ser distribuídas para outros OSDs diferentes de modo mais homogêneo possível, e uma mesma exigência é verdadeira para um componente físico de nível mais alto.

[0066] Conforme mostrado na Figura 3, em uma modalidade específica da presente invenção, que o módulo de MDC gera informações de visualização de agrupamento, pode incluir especificamente: o MDC gera as informações de visualização de agrupamento de acordo com informações de configuração de agrupamento entregues por um administrador e a situação de particionamento. Especificamente, as informações de visualização de agrupamento incluem visualizações de agrupamento em três dimensões: uma visualização de OSD (OSD view), uma visualização de IO (IO View) e uma visualização de partição (partition view).

[0067] A visualização de OSD inclui informações de situação de um nó de OSD em um agrupamento. Em um modo de implantação específico, a visualização de OSD pode incluir um ID de um nó de OSD e as informações de situação do nó de OSD, em que o ID de OSD é um marcador ou número de OSD. Em uma modalidade da presente invenção mostrada na Figura 4, uma situação de OSD pode ser especificamente classificada em um estado "ASCENDENTE (UP)" e um estado "DESCENDENTE (DOWN)" acordo com o fato de se o OSD é defeituoso, e classificado em um estado "FORA (OUT)" e um estado "DENTRO (IN)" de acordo com o fato de se o OSD sai de um agrupamento. Conforme mostrado na Figura 4, uma transição de situação específica inclui: após um failback, um nó de OSD é inicializado ou reiniciado e, então, transita de um

estado "DENTRO (IN)" e "DESCENDENTE (DOWN)" para um estado "DENTRO (IN)" e "ASCENDENTE (UP)". Quando uma falha de um OSD dura por um limiar específico (por exemplo, mais que 5 minutos), o nó de OSD é expulso do agrupamento e, consequentemente, o nó de OSD transita de um estado "DENTRO (IN)" e "DESCENDENTE (DOWN)" para um estado "FORA (OUT)" e "DESCENDENTE (DOWN)". Em uma modalidade específica da presente invenção, a visualização de OSD pode incluir adicionalmente informações de versão de visualização de OSD, tais como um número da versão de visualização de OSD, uma visualização de ID de OSD, ou quaisquer outras informações que marquem uma versão de visualização.

[0068] A visualização de IO inclui um identificador que identifica um nó de OSD primário de um grupo de partição. Em um modo de implantação específico, a visualização de IO pode incluir um ID de grupo de partição e um identificador de um nó de OSD primário de um grupo de partição que correspondem ao ID de grupo de partição. Cada visualização de IO tem informações de versão de visualização de IO que identificam a visualização de IO, em que as informações de versão de visualização de IO podem ser um ID de visualização de IO (que também pode ser chamado de um número de versão de visualização de IO), que é usado para identificar uma versão da visualização de IO, de modo a auxiliar diferentes módulos a comparar as versões de visualização de IO. Em uma modalidade específica, as informações de versão de visualização de IO podem ser incluídas na visualização de IO, ou podem ser excluídas da visualização de IO.

[0069] A visualização de partição inclui informações a respeito de um OSD em que uma partição em um grupo de partição está localizada. Em um modo de implantação específico, a visualização de partição pode incluir um ID de grupo de partição, um OSD no qual cada partição em um grupo de partição que correspondem ao ID de grupo de partição está localizada e uma identidade primária/secundária do OSD, e uma situação de partição que corresponde ao OSD de cada partição. A visualização de partição inclui informações a respeito de um nó de OSD no qual uma partição primária está localizada (tal como um nó de ID de OSD, uma identidade primária/secundária do nó de OSD e uma situação de partição que corresponde ao nó de OSD da partição primária) e informações a respeito de um nó de OSD no qual uma partição secundária

(podem haver uma ou mais partições secundárias) que corresponde à partição primária está localizada (tal como um nó de ID de OSD, uma identidade primária/secundária do nó de OSD e uma situação de partição da partição secundária que corresponde ao OSD). Em uma modalidade específica, a situação de partição pode ser classificada em dois tipos: "Consistente" e "inconsistente", em que "consistente" indica que os dados em uma partição secundária são consistentes com aqueles em uma partição primária, e "inconsistente" indica que os dados em uma partição secundária podem ser inconsistentes em relação aqueles em uma partição primária. Cada visualização de partição tem informações de versão de visualização de partição que identificam a visualização de partição, em que as informações de versão de visualização de partição podem ser um ID de visualização de partição (que também pode ser chamado de um número de versão de visualização de partição), de modo que os módulos comparem as visualizações. Em uma modalidade específica, as informações de versão de visualização de partição podem ser incluídas na visualização de partição, ou podem ser excluídas da visualização de partição. Em uma modalidade específica, devido ao fato de que a visualização de IO é um subconjunto da visualização de partição, ou seja, a visualização de partição inclui as informações de visualização de IO, a visualização de partição pode incluir adicionalmente as informações de versão de visualização de IO.

[0070] O MDC é adicionalmente adaptado para: manter, gerenciar e atualizar a visualização de agrupamento; atualizar a visualização de agrupamento de acordo com um nó de situação de OSD, tal como uma falha, um failback, sair de um agrupamento após uma falha, se reunir a um agrupamento após um failback, e se unir a um agrupamento novo; e notificar um módulo relacionado de atualização de visualização de agrupamento, de modo que o módulo relacionado processe, de acordo com a visualização de agrupamento atualizada, a replicação de dados que correspondem a uma solicitação de IO correspondente.

[0071] Em um modo de implantação específico, para se reduzir a interação e economizar recursos de gerenciamento e armazenamento, a visualização de OSD pode existir apenas no MDC, a visualização de partição pode existir apenas no módulo de MDC e o nó de OSD primário, e a visualização

de IO existe no módulo de MDC, no módulo de roteamento de IO, no nó de OSD primário e no nó de OSD secundário. O módulo de MDC envia a visualização de partição apenas a um nó de OSD primário no qual uma partição na visualização de partição está localizada, ou instrui apenas um nó de OSD primário no qual uma partição na visualização de partição está localizada para atualizar uma visualização de partição local; e envia a visualização de IO que constitui uma parte da visualização de partição (ou seja, a visualização de IO pode ser considerada como uma subvisualização da visualização de partição) para o módulo de roteamento de IO, o nó de OSD primário e o nó de OSD secundário, ou instrui um módulo correspondente a atualizar uma visualização de IO armazenada localmente. Para um processo de implantação específico, referir-se aos procedimentos específicos a seguir e a um procedimento em que um OSD se une a um agrupamento. Em uma aplicação real, o módulo de MDC pode definir visualizações de agrupamento em diferentes formas de acordo com informações de configuração ou uma norma específica e de acordo com funções básicas das visualizações de agrupamento, que não é limitada nessa modalidade da presente invenção.

[0072] Em uma modalidade específica da presente invenção, o módulo de roteamento de IO é adaptado principalmente para implantar uma função de roteamento de solicitação de IO. O módulo de roteamento de IO adquire as visualizações de IO de todas as partições no agrupamento a partir do módulo de MDC e armazena em cache as visualizações de IO. Quando a solicitação de IO de serviço chega no módulo de roteamento de IO, o módulo de roteamento de IO obtém, por meio de cálculo (um algoritmo de hash ou outro algoritmo pode ser usado em um método de cálculo) através do uso de uma chave na solicitação de IO, um grupo de partição ao qual o IO pertence; então, procura a as visualizações de IO armazenadas localmente para encontrar um nó de OSD primário que corresponde ao grupo de partição, e envia a solicitação de IO para o nó de OSD primário. O módulo de roteamento de IO processa uma notificação de atualização de visualização de IO recebida a partir do módulo de MDC, em que a notificação de atualização pode incluir a visualização de IO atualizada ou informações de indicação de atualização correspondentes, que indicam, por exemplo, o conteúdo que deve ser atualizado; atualiza uma visualização de IO armazenada localmente de acordo com a notificação de

atualização, e roteia a solicitação de IO de acordo com a visualização de IO de armazenamento local atualizada. Para um processo de implantação específico, referir-se aos procedimentos específicos a seguir.

[0073] Em uma modalidade específica da presente invenção, que o nó de OSD processa a solicitação de IO de acordo com a visualização de agrupamento para executar uma operação de IO inclui especificamente o seguinte:

[0074] Quando o nó de OSD serve como um nó de OSD primário, o nó de OSD primário é principalmente adaptado para: receber a solicitação de IO enviada pelo módulo de roteamento de IO, executar a solicitação de IO e enviar uma solicitação de replicação para um nó de OSD secundário correspondente, de modo a executar o armazenamento e replicação de dados de IO. O nó de OSD primário recebe uma visualização de partição de uma partição no nó de OSD primário a partir do módulo de MDC e armazena a visualização de partição. O nó de OSD primário processa a replicação da solicitação de IO de acordo com a visualização de partição. O nó de OSD primário recebe adicionalmente, a partir do módulo de MDC, uma notificação de atualização em relação à visualização de partição, atualiza a visualização de partição armazenada localmente de acordo com a notificação de atualização e processa, de acordo com a visualização de partição atualizada, a replicação de dados que correspondem à solicitação de IO, em que a notificação de atualização pode incluir a visualização de partição atualizada ou informações de atualização correspondentes, de modo que o nó de OSD atualize a visualização de partição armazenada localmente e a visualização de IO de acordo com a visualização de partição atualizada ou as informações de atualização. Quando o nó de OSD serve como um nó de OSD secundário, o nó de OSD secundário é adaptado para: receber uma solicitação de replicação de um nó de OSD primário, e realizar a replicação e cópia de segurança de dados de acordo com a solicitação de replicação; receber uma visualização de IO de uma partição à qual os dados no nó de OSD secundário pertencem a partir do módulo de MDC, e armazenar a visualização de IO e processar, de acordo com a visualização de IO, a replicação de dados que correspondem à solicitação de IO; e receber adicionalmente, a partir do módulo de MDC, uma notificação de atualização em relação à visualização de IO, atualizar a visualização de IO armazenada

localmente de acordo com a notificação de atualização e processar, de acordo com a visualização de IO atualizada, a replicação dos dados que correspondem à solicitação de IO. Para um processo de implantação específico, referir-se aos procedimentos específicos a seguir.

[0075] Em uma modalidade específica da presente invenção, o sistema de armazenamento e replicação distribuído (mostrado na Figura 2A, Figura 2B e Figura 2C) na modalidade anterior pode ser implantado com base em um sistema mostrado na Figura 5. Conforme mostrado na Figura 5, o sistema pode incluir uma ou mais memórias 502, uma ou mais interfaces de comunicações 504 e um ou mais processadores 506 ou outra rede de interação de dados (que é usada para a interação entre múltiplos processadores e memórias, e não é mostrada na figura).

[0076] A memória 502 pode ser uma memória de diversos tipos, tal como uma memória somente de leitura (Read Only Memory, ROM), um dispositivo de armazenamento estático, um dispositivo de armazenamento dinâmico, ou uma memória de acesso aleatório (Random Access memória, RAM). A memória 502 pode armazenar um sistema operacional e uma instrução e dados de aplicação de ouro programa de aplicativo, em que a instrução inclui uma instrução usada para se executar as funções de um módulo de MDC, e um módulo de roteamento de IO e um nó de OSD em diversas modalidades da presente invenção.

[0077] A instrução armazenada na memória 502 é rodada e executada pelo processador 506.

[0078] A interface de comunicações 504 é adaptada para implantar a comunicação entre a memória 502 e o processador 506, a comunicação entre os processadores, a comunicação entre as memórias e a comunicação entre o sistema e outro dispositivo ou uma rede de comunicações.

[0079] O processador 506 pode ser uma unidade de processamento central de propósito geral (Central Processing Unit, CPU), um microprocessador, um circuito integrado para aplicação específica (Application Specific Integrated Circuit, ASIC), ou uma ou mais circuitos integrados, e é adaptado para executar um programa relacionado, de modo a executar os procedimentos de interação entre o módulo de MDC, o módulo de roteamento de IO e o nó de OSD descritos e as funções implantadas em várias modalidades

da presente invenção.

[0080] Para facilitar a compreensão e evitar descrições repetidas desnecessárias, as modalidades específicas a seguir são usadas para descrever como os procedimentos de interação entre o módulo de MDC, o módulo de roteamento de IO e o nó de OSD descritos aqui e as funções implantadas nas modalidades da presente invenção são executados no sistema mostrado na Figura 5. Com base em todas as modalidades da presente invenção, um indivíduo versado na técnica pode compreender que o sistema mostrado na Figura 5 pode ser adaptado para implantar casos descritos em várias outras modalidades, todas as quais estão dentro do escopo registrado e revelado da presente invenção.

MODALIDADE 1

[0081] O processador é adaptado para: ser conectado à memória e ler a instrução na memória, sendo que a instrução inclui a instrução usada para executar as funções do módulo de MDC, do módulo de roteamento de IO, e do nó de OSD; e permitir, de acordo com a instrução, que o processador executar as seguintes operações:

capacitar o módulo de MDC para: determinar que um nó de OSD no sistema é um nó de OSD defeituoso, determinar uma partição no nó de OSD defeituoso, atualizar uma visualização de partição de um grupo de partição que inclui a partição no nó de OSD defeituoso, e enviar uma notificação de atualização para um nó de OSD primário no qual o grupo de partição na visualização de partição atualizada está localizado, de modo que o nó de OSD primário processe, de acordo com a visualização de partição atualizada, a replicação de dados que correspondem a uma solicitação de IO.

[0082] Deve-se compreende que, na modalidade anterior, as funções do módulo de MDC supracitado, do módulo de roteamento de IO e do nó de OSD podem ser implantadas por um hospedeiro. Nesse caso, a instrução para implantar as funções do módulo de MDC supracitado, do módulo de roteamento de IO e do nó de OSD pode existir em uma memória do hospedeiro, e um processador do hospedeiro lê a instrução para implantar as funções do módulo de MDC, do módulo de roteamento de IO e do nó de OSD anterior a partir da memória. Em outra modalidade, as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD podem ser implantadas por

múltiplos hospedeiros de uma maneira interativa. Nesse caso, o módulo de MDC anterior, o módulo de roteamento de IO e o nó de OSD são armazenados em memórias dos diferentes hospedeiros de uma maneira distribuída. Por exemplo, um processador de um hospedeiro 1 executa uma função do módulo de MDC anterior, um processador de um hospedeiro 2 executa uma função do nó de OSD primário, e um hospedeiro 3 executa uma função do módulo de roteamento de IO.

MODALIDADE 2

[0083] O processador é adaptado para: ser conectado à memória e ler a instrução na memória, sendo que a instrução inclui a instrução usada para executar as funções do módulo de MDC, do módulo de roteamento de IO, e do nó de OSD; e permitir, de acordo com a instrução, que o processador executar as seguintes operações:

capacitar o módulo de roteamento de IO para: receber uma solicitação de IO, em que a solicitação de IO inclui uma chave; determinar, de acordo com a chave, um grupo de partição ao qual os dados que correspondem à solicitação de IO pertencem, e determinar um nó de OSD primário do grupo de partição ao qual os dados pertencem; adicionar informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem para a solicitação de IO, e enviar, para o nó de OSD primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO;

capacitar o nó de OSD primário para: receber a solicitação de IO; executar a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO armazenada localmente; gerar uma solicitação de replicação que porta as informações de versão de visualização de IO; e enviar a solicitação de replicação para um nó de OSD secundário da partição à qual os dados pertencem; e

capacitar o nó de OSD secundário para: receber a solicitação de replicação, e executar a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO armazenada localmente no nó de OSD secundário, de modo que os dados que correspondem à solicitação de IO no nó de OSD

secundário permaneçam consistentes com os dados que correspondem à solicitação de IO no nó de OSD primário.

[0084] De modo semelhante, na modalidade anterior, as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD podem ser implantadas por um hospedeiro. Nesse caso, a instrução para implantar as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD pode existir em uma memória do hospedeiro, e um processador do hospedeiro lê a instrução para implantar as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD a partir da memória. Em outra modalidade, as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD podem ser implantadas por múltiplos hospedeiros de uma maneira interativa. Nesse caso, o módulo de MDC anterior, o módulo de roteamento de IO e o nó de OSD são armazenados em memórias dos diferentes hospedeiros de uma maneira distribuída. Por exemplo, um processador de um hospedeiro 1 executa uma função do módulo de roteamento de IO anterior, um processador de um hospedeiro 2 executa uma função do nó de OSD primário, um processador de um hospedeiro 3 executa uma função do nó de OSD secundário e um processador de um hospedeiro 4 executa uma função do módulo de MDC.

MODALIDADE 3

[0085] O processador é adaptado para: ser conectado à memória e ler a instrução na memória, sendo que a instrução inclui a instrução usada para executar as funções do módulo de MDC, do módulo de roteamento de IO, e do nó de OSD; e permitir, de acordo com a instrução, que o processador executar as seguintes operações:

permitir que o nó de OSD: envie uma solicitação de consulta para o módulo de MDC depois de um failback para solicitar uma visualização de IO de um grupo de partição que inclui uma partição no nó de OSD, sendo que o nó de OSD é chamado de um nó de OSD de failback, e a solicitação de consulta carrega um identificador de OSD do nó de OSD de failback; receba a visualização de IO devolvida pelo MDC; inicie uma solicitação de recuperação de dados para um OSD primário na visualização de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD de failback durante falha; receba os dados que são atualizados durante a falha e que são enviados pelo

OSD primário; e processar a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC depois de o nó de OSD de failback completar a recuperação de dados;

permitir que o módulo de MDC: receba a solicitação de consulta do nó de OSD de failback, devolva a visualização de IO para o nó de OSD de failback de acordo com o identificador de OSD na solicitação de consulta, e atualize a visualização de partição do grupo de partição depois de o nó de OSD de failback completar a recuperação de dados; e

permitir que o nó de OSD primário: receba a solicitação de recuperação de dados do nó de OSD de failback, envie os dados atualizados durante a falha para o nó de OSD de failback, e processe, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC depois de o nó de OSD de failback completar a recuperação de dados, a replicação de dados que correspondem à solicitação de IO.

[0086] De modo similar, na modalidade anterior, as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD podem ser implantados por um hospedeiro. Nesse caso, a instrução para implantar as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD pode existir em uma memória do hospedeiro, e um processador do hospedeiro lê a instrução para implantar as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD a partir da memória. Em outra modalidade, as funções do módulo de MDC anterior, do módulo de roteamento de IO e do nó de OSD podem ser implantadas por múltiplos hospedeiros de uma maneira interativa. Nesse caso, o módulo de MDC anterior, o módulo de roteamento de IO e o nó de OSD são armazenados em memórias dos diferentes hospedeiros de uma maneira distribuída. Por exemplo, um processador de um hospedeiro 1 executa uma função do nó de OSD de failback anterior, um processador de um hospedeiro 2 executa uma função do nó de OSD primário, um processador de um hospedeiro 3 executa uma função do módulo de MDC, e um processador de um hospedeiro 4 executa uma função do módulo de roteamento de IO.

[0087] A seguir são usadas múltiplas modalidades de procedimento específicas para descrever adicionalmente em detalhes a conexão e a interação,

as funções específicas, e similares do módulo de MDC 202, do módulo de roteamento de IO 204, e do nó de OSD 206 no sistema de controle de replicação e armazenamento distribuído anterior. Essas modalidades de procedimento específicas incluem: um procedimento de geração e aquisição inicializada de uma visualização de agrupamento, um procedimento de processamento de solicitação de IO, um procedimento de processamento de falha de OSD, um procedimento de failback de nó de OSD, um procedimento de recuperação de dados, um procedimento em que um nó de OSD existe a partir de um agrupamento depois de uma falha, e um procedimento em que um novo nó de OSD une um agrupamento, os quais são descritos um a um em detalhes a seguir.

[0088] Deve-se observar que nem todas as etapas ou funções nas modalidades a seguir são necessários; uma sequência de etapas é meramente para facilitar a descrição, e não é particularmente limitada a não ser que seja necessariamente exigida de acordo com o princípio das soluções da presente invenção; adicionalmente, a maneiras de implantação específicas nas etapas são meramente usadas como exemplos ilustrativos, e não constituem limitação particular ao escopo de proteção da presente invenção. Com base no relatório descritivo inteiro da presente invenção, um indivíduo versado na técnica pode fazer adição ou exclusão correspondentes ou variação ou substituição não criativa das etapas anteriores de acordo com uma situação real, e combinar diferentes maneiras de implantação nas etapas com maneiras de implantação em outras etapas para formar diferentes modalidades, todas as modalidades estão dentro do escopo gravado e revelado da presente invenção.

[0089] O procedimento de geração e aquisição inicializadas de uma visualização de agrupamento:

[0090] A Figura 6 mostra uma modalidade de geração e aquisição inicializadas de uma visualização de agrupamento de acordo com a presente invenção. Nessa modalidade, um módulo de MDC gera uma visualização de agrupamento inicial de acordo com as informações de configuração de agrupamento entregues por um administrador. Um módulo de roteamento de IO e um nó de OSD consultam o módulo de MDC para a visualização durante a inicialização. Essa modalidade é executada pelo módulo de MDC, o módulo de roteamento de IO, e o nó de OSD mencionado nas modalidades descritas nas Figura 2A a Figura 2C e na Figura 5.

[0091] Durante a inicialização do sistema, o módulo de MDC é primeiro iniciado e, depois, a aquisição, pelo módulo de roteamento de IO e o nó de OSD, da visualização do MDC é iniciada. Um processo específico inclui as seguintes etapas:

[0092] 602. Um usuário ou o administrador entrega as informações de configuração de agrupamento para o módulo de MDC, sendo que as informações de configuração de agrupamento podem incluir os parâmetros de configuração de sistema tais como as informações de topologia de agrupamento, uma quantidade de partição, e uma quantidade de réplica e as informações de topologia de agrupamento incluem principalmente uma quantidade de servidores e seus endereços de IP, informações de rack, uma quantidade de nós de OSD em cada servidor e informações sobre recursos de armazenamento físicos gerenciados pelos nós de OSD (por exemplo, informações de disco local que correspondem aos nós de OSD) e similares.

[0093] 604. O MDC gera a visualização inicial de agrupamento de acordo com a configuração de agrupamento entregue, sendo que os três tipos de visualizações de agrupamento (uma visualização de OSD, uma visualização de partição e uma visualização de IO) já foram descritas acima; o MDC gera uma visualização de OSD de acordo com as informações de OSD configuradas, gera uma visualização de partição com o uso de um algoritmo de alocação de partição, a quantidade de partição configurada, quantidade de réplica, e quantidade de nó de OSD, e não precisa gerar adicionalmente uma visualização de IO, que é um subconjunto da visualização de partição, sendo que quando a visualização de partição está sendo gerada, o equilíbrio de alocação de partição (uma quantidade de partições em cada nó de OSD se mantém a mesma tanto quanto possível) e a segurança (nós de OSD nos quais as réplicas de partição existem estão em diferentes servidores ou diferentes racks) geralmente precisam ser considerados.

[0094] 606. Começar a inicialização do módulo de roteamento de IO, e consultar o módulo de MDC pela visualização de IO, sendo que ao ser iniciado, o módulo de roteamento de IO precisa adquirir uma visualização do MDC relacionada de modo a funcionar normalmente.

[0095] 608. Começar a inicialização do nó de OSD, e consultar o módulo de MDC pela visualização de partição e pela visualização de IO, sendo

que o nó de OSD precisa adquirir uma visualização de partição de um grupo de partição que inclui uma partição primária distribuída no nó de OSD e uma visualização de IO de um grupo de partição que inclui uma partição secundária distribuída no nó de OSD.

[0096] 610. O módulo de MDC devolve visualizações de IO de todos os grupos de partição para o módulo de roteamento de IO.

[0097] 612. O módulo de MDC devolve, para o nó de OSD, a visualização de partição do grupo de partição que inclui a partição primária distribuída no OSD e a visualização de IO do grupo de partição que inclui a partição secundária distribuída no nó de OSD.

[0098] Procedimento de processamento de solicitação de IO:

[0099] A Figura 7A e a Figura 7B mostram uma modalidade de um procedimento de solicitação de IO de acordo com a presente invenção. Essa modalidade é executada pelo módulo de roteamento de IO e pelo nó de OSD mencionado nas modalidades descritas na Figura 2A à Figura2 C e na Figura 5. Para facilitar a compreensão, um grupo de partição (uma partição X) é usado como um exemplo para descrição nessa modalidade, sendo que a partição X pode ser qualquer partição gerenciada e mantida por uma camada de protocolo de replicação distribuída nessa modalidade da presente invenção, um nó de OSD primário no qual a partição X está localizada é abreviado como um nó de OSD primário de Partição X, um nó de OSD1 secundário no qual uma partição secundária na Partição X está localizada é abreviado como um nó de OSD1 secundário de Partição X, e um nó de OSD2 secundário no qual uma partição secundária na Partição X está localizada é abreviado como um nó de OSD2 secundário de Partição X. A partição X é usada como um exemplo na descrição específica a seguir. Um procedimento de operação de IO (por exemplo, uma operação de escrita ou uma operação de modificação) nessa modalidade especificamente inclui as seguintes etapas:

[0100] 702. O módulo de roteamento de IO recebe uma solicitação de IO enviada por um hospedeiro (por exemplo, um servidor mostrado na Figura 2C no qual o módulo de roteamento de IO está localizado).

[0101] 704. O módulo de roteamento de IO obtém, de acordo com a solicitação de IO recebida, uma partição de dados (que pode também ser chamada de dados de IO) que corresponde à solicitação de IO, e adquire um nó

de OSD primário de um grupo de partição que inclui a partição.

[0102] Em uma maneira de implantação específica, o módulo de roteamento de IO pode obter, por meio de cálculo com o uso de um algoritmo hash e de acordo com uma chave carregada na solicitação de IO, um ID de partição do grupo de partição que inclui os dados que correspondem à solicitação de IO e, depois, procura por uma visualização de IO com o uso do ID de partição, de modo a adquirir o nó de OSD primário do grupo de partição. Conforme descrito acima, o grupo de partição que corresponde ao ID de partição no presente documento é a Partição X nessa modalidade. A chave é um dígito ou sequência de caracteres definida em um serviço de camada superior, e é usada para identificar um bloco de dados.

[0103] 706. O módulo de roteamento de IO envia a solicitação de IO para o nó de OSD primário do grupo de partição, sendo que a solicitação carrega informações de versão de visualização de IO (por exemplo, um ID de visualização de IO), informações de chave de IO e os dados de IO.

[0104] De acordo com a descrição anterior, o grupo de partição nessa modalidade é a Partição X e, consequentemente, a solicitação de IO é enviada para o nó de OSD primário da partição X, isto é, o nó de OSD primário de partição X. Em uma maneira de implantação específica, um ID de visualização de IO pode ser chamado também de um número de versão de visualização de IO e o número de versão de visualização de IO é usado principalmente para identificar uma versão de visualização e cresce de modo monotônico, sendo que um pequeno ID de visualização de IO indica que uma versão de visualização mantida pelo nó de OSD primário é uma versão desatualizada, e uma exigência para garantir consistência é que as visualizações vistas por todos os módulos no procedimento de processamento de IO sejam consistentes. As informações de chave de IO podem incluir uma chave, um desvio e um comprimento, sendo que o desvio indica um desvio dos dados de IO com relação a uma posição de início em um bloco de dados identificado pela chave, e o comprimento indica um comprimento dos dados de IO.

[0105] 708. Gerar um ID de sequência (Seq) (que pode ser chamada também de um identificador de sequência) para a solicitação de IO depois de ser determinado, de acordo com as informações de versão de visualização de IO, que uma versão de visualização carreada na solicitação de

IO é consistente com uma versão de visualização armazenada localmente, sendo que especificamente, depois de determinar que o ID de visualização de IO carregado na solicitação de IO é consistente com um ID de visualização de IO armazenada localmente, o nó de OSD primário gera um identificador de sequência para a solicitação de IO.

[0106] Em uma maneira de implantação específica, o ID de Seq inclui o número de versão de visualização e um número de sequência (Seq NO). O número de versão de visualização aumenta de modo monotônico à medida que a visualização de IO muda, e o Seq NO indica um número de série de uma operação de modificação (por exemplo, escrever e excluir) nos dados que correspondem a uma partição na visualização de IO dentro de uma versão de visualização de IO. Depois de a visualização de IO mudar, o Seq NO no ID de Seq começa a aumentar novamente do 0. O nó de OSD primário de partição X comparar se o ID de visualização de IO carregado na solicitação de IO é consistente com o ID de visualização de IO localmente armazenado pode ser especificamente: o nó de OSD primário de partição X primeiro compara os IDs de visualização de IO, sendo que um ID de visualização de IO maior indica um ID de Seq maior; e se os IDs de visualização de IO são iguais, então, compara os Seq NOs, sendo que um Seq NO maior indica um ID de Seq maior, e apenas quando tanto os IDs de visualização de IO quanto os Seq NOs são os mesmos, isso indica que os Seq IDs são consistentes.

[0107] 710. Se a visualização de IO local é maior, devolver um erro para o módulo de roteamento de IO; ou se a visualização de IO local é menor, adicionar a solicitação de IO a uma fila de cache, e consultar o módulo de MDC por uma visualização de partição.

[0108] Em uma maneira de implantação específica, depois de determinar que o ID de visualização de IO na solicitação de IO é menor do que o ID de visualização de IO armazenado localmente, o nó de OSD primário de partição X devolve o erro para o módulo de roteamento de IO, e o módulo de roteamento de IO consulta o módulo de MDC para a visualização de IO do grupo de partição, e reenvia a solicitação de IO depois de obter um ID de visualização de IO atualizado; ou depois de determinar que o ID de visualização de IO na solicitação de IO é maior do que o ID de visualização de IO localmente armazenado, o nó de OSD primário de partição X adiciona a solicitação de IO à

fila de cache, e consulta o módulo de MDC por um ID de visualização de IO da visualização de IO do grupo de partição, de modo a executar a solicitação de IO depois de determinar que o ID de visualização localmente armazenado de IO é consistente com o ID de visualização de IO na solicitação de IO.

[0109] 712. Gravar uma entrada, sendo que a entrada inclui um tipo de operação, um ID de grupo de partição, um ID de Seq e uma chave.

[0110] Em uma maneira de implantação específica, o tipo de operação pode incluir escrever, excluir ou similares. Para uma operação de escrita, a entrada pode incluir adicionalmente o deslocamento e comprimento anteriores. Adicionalmente, para vários tipos de operações, a entrada pode incluir adicionalmente informações de situação, que são usadas para descrever se a operação é bem-sucedida. Em geral, todas as operações de modificação (tal como operações de escrita e exclusão) em um mesmo grupo de partição são consecutivamente numeradas.

[0111] 714. Executar a solicitação de IO, e consultar uma visualização de partição do grupo de partição para obter informações sobre um nó de OSD secundário do grupo de partição.

[0112] Em uma maneira de implantação específica, se a execução da solicitação de IO é escrever a solicitação de IO, os dados de IO são escritos em recursos de armazenamento físicos (por exemplo, uma camada de cache ou uma camada de persistência mostradas na Figura 2C, tal como um disco magnético, ou os recursos de armazenamento físicos externos SAN mencionados anteriormente) gerenciados pelo nó de OSD primário de partição X; e se a solicitação de IO é uma solicitação de exclusão, dados correspondentes nos recursos de armazenamento físicos locais gerenciados pelo nó de OSD primário de partição X são excluídos. Em uma modalidade específica, o nó de OSD primário de partição X gera adicionalmente uma solicitação de replicação. Em uma implantação específica, a solicitação de replicação pode ser uma solicitação de replicação gerada montando-se separadamente uma parte de controle da solicitação de IO, de modo que dados que correspondem à solicitação de IO no nó de OSD secundário da partição X sejam consistentes com dados que correspondem à solicitação de IO no nó de OSD primário de partição X.

[0113] 716(1)/(2). Enviar separadamente uma solicitação de

replicação para o nó de OSD1 secundário de Partição X e o nó de OSD2 secundário de Partição X, sendo que a solicitação de replicação carrega o ID de Seq.

[0114] Em uma maneira de implantação específica, a solicitação de replicação pode incluir adicionalmente as informações tais como a chave, o desvio, o comprimento e os dados de IO na solicitação original. Quando a solicitação de replicação é uma solicitação de replicação de escrita, a solicitação de replicação carrega a chave, o desvio, o comprimento e os dados de IO; e quando a solicitação de replicação é uma solicitação de replicação de exclusão, a solicitação de replicação carrega apenas a chave.

[0115] 718. Processar a solicitação de replicação depois de ser determinado que um ID de visualização de IO na solicitação de replicação é consistente com um ID de visualização localmente armazenado de IO.

[0116] Em uma maneira de implantação específica, adicionalmente à etapa anterior 718, o procedimento pode incluir adicionalmente as seguintes etapas:

[0117] depois de determinar que o ID de visualização de IO na solicitação de replicação é menor do que o ID de visualização localmente armazenado de IO, o nó de OSD1 secundário de Partição X ou o OSD2 secundário de Partição X devolve um erro para o nó de OSD primário de partição X, e o nó de OSD primário de partição X consulta o módulo de MDC pela visualização de IO do grupo de partição, e reenvia a solicitação de replicação depois de obter o ID de visualização de IO atualizado; ou depois de determinar que o ID de visualização de IO na solicitação de replicação é maior do que o ID de visualização localmente armazenado de IO, o nó de OSD1 secundário de Partição X ou o nó de OSD2 secundário de Partição X adiciona a solicitação de replicação à fila de cache, e consulta o módulo de MDC pelo número de versão de visualização de IO do grupo de partição, de modo a executar a solicitação de replicação depois de determinar que o ID de visualização localmente armazenado de IO é consistente com o ID de visualização de IO na solicitação de IO.

[0118] No protocolo de compromisso de duas fases existente, se um participante rejeita uma proposta, um procedimento de IO inteiro precisa ser revertido, o que causa uma sobrecarga bem grande. Contudo, nessa

modalidade, se um nó de OSD secundário rejeita uma solicitação, uma última visualização, primeiro, é consultada e, depois, o processamento continua, e uma reversão não precisa ser realizada, o que, desse modo, aprimora a tolerância à falha e disponibilidade de um sistema inteiro.

[0119] No protocolo de comprometimento de duas fases existente, em um processo de IO normal, a interação de mensagem precisa ser realizada entre um coordenador e um participante duas vezes. Contudo, nessa modalidade, um processo de IO precisa ser realizado apenas uma vez, e a interação de mensagem é realizada entre um nó primário e um nó secundário apenas uma vez, o que reduz um atraso de IO causado pela interação de mensagem e aprimora a eficiência e desempenho do sistema inteiro.

[0120] Em uma maneira de implantação específica, a solicitação de replicação pode carregar adicionalmente um ID de Seq em uma solicitação de replicação prévia enviada pelo nó de OSD primário de partição X da partição X, e o procedimento de IO pode incluir adicionalmente a etapa 720.

[0121] 720. Compare a Seq ID que é uma solicitação de replicação anterior e que é carregada na solicitação de replicação com um ID de Seq local maior, e se o ID de Seq for maior do que o maior ID de Seq local, exigir que o nó de OSD primário de Partição X envie uma solicitação perdida, ou se o ID de Seq for consistente com o maior ID de Seq, continuar o processamento em etapas subsequentes.

[0122] A solicitação de replicação carrega o ID de Seq que está na solicitação de replicação anterior, o que pode evitar a perda de dados causada por uma mudança no número de versão de visualização de IO. Adicionalmente, dados ausentes são enviados, o que pode garantir que as sequências em que um nó de OSD primário e um nó de OSD secundário executam todos os IO sejam consistentes, e aprimorar adicionalmente a consistência da cópia de segurança de dados.

[0123] Em uma maneira de implantação específica, se o ID de Seq que está na solicitação de replicação anterior e é carregado na solicitação de replicação é menor do que o maior ID de Seq local, um erro é devolvido para o nó de OSD primário de partição X, e o nó de OSD primário de partição X reenvia a solicitação de replicação; ou um ID de Seq é determinado por meio de consulta adicional e, depois de o ID de Seq ser determinado, um ID de Seq atualizado é

adquirido em depois, o processamento continua, em vez da devolução direta de um erro. Nos dois casos, o processamento não precisa ser terminado para uma reversão, o que aprimora adicionalmente a tolerância à falha e a disponibilidade do sistema e o desempenho do sistema inteiro.

[0124] Quando as operações de IO precisam ser executadas tanto pelo nó de OSD primário quanto pelo nó de OSD secundário, exige-se garantir que as sequências em que o nó de OSD primário e o nó de OSD secundário executam todas as operações de IO sejam consistentes, o que é uma chave para garantir a consistência de múltiplas réplicas de partição.

[0125] 722. Gravar uma entrada, e executar a operação de replicação.

[0126] Em uma maneira de implantação específica, a entrada inclui um tipo de operação, um ID de partição, um ID de Seq, e uma chave, sendo que o tipo de operação pode incluir escrever, excluir ou similares. Para uma operação de escrita, a entrada pode incluir adicionalmente o desvio e o comprimento anteriores. Adicionalmente, para vários tipos de operações, a entrada pode incluir adicionalmente informações de situação, que são usadas para descrever se a operação é bem-sucedida. A execução da solicitação de replicação inclui especificamente: quando a solicitação de replicação é uma solicitação de replicação de escrita, a solicitação de replicação carrega a chave, o desvio, o comprimento e os dados de IO; e quando a solicitação de replicação é uma solicitação de replicação de exclusão, a solicitação de replicação carrega apenas a chave.

[0127] 724/726 O nó de OSD1 secundário de Partição X e o nó de OSD2 secundário de Partição X enviam separadamente uma mensagem de reposta de sucesso de solicitação para o nó de OSD primário de Partição X.

[0128] 728. O nó de OSD primário de Partição X envia a resposta de sucesso de solicitação de resposta para o módulo de roteamento de IO.

[0129] Com base na modalidade de solicitação de IO anterior, considerando-se que um processo de processamento da solicitação de IO, do nó de OSD primário de partição X ou do nó de OSD secundário de Partição X se torna defeituoso (por exemplo, depois de a solicitação de IO chegar ao nó de OSD primário de partição X, o nó de OSD primário de partição X se torna defeituoso), ou o nó de OSD secundário de Partição X se torna defeituoso, um

novo nó de OSD une o sistema à medida que um nó de OSD secundário da partição X, ou similares, nesses casos, o procedimento de processamento de solicitação de IO na modalidade anterior pode incluir adicionalmente um processo de processamento descrito na modalidade a seguir.

[0130] Quando o nó de OSD primário de partição X se torna defeituoso depois de a solicitação de IO chegar ao nó de OSD primário de partição X, o procedimento de processamento de IO inclui o seguinte:

depois de o nó de OSD primário de partição X se tornar defeituoso, o módulo de MDC no sistema define, ao detectar, no processo de processamento da solicitação de IO, que o nó de OSD primário de partição X se torna defeituoso, o nó de OSD primário de partição X na visualização de partição do grupo de partição como um novo nó de OSD secundário de Partição X, e marca uma situação de partição do novo nó de OSD secundário de Partição X como inconsistente; define o nó de OSD1 secundário de Partição X na visualização de partição do grupo de partição como um novo nó de OSD primário de partição X, e envia a visualização de partição atualizada de grupo de partição para o novo nó de OSD primário de partição X; e define o nó de OSD1 secundário de Partição X na visualização de IO do grupo de partição como um novo nó de OSD primário de partição X, e envia a visualização de IO do grupo de partição atualizada para o módulo de roteamento de IO;

o módulo de roteamento de IO recebe a visualização de IO atualizada que é do grupo de partição e que é enviada pelo módulo de MDC, e envia a solicitação de IO para o novo nó de OSD primário de partição X de acordo com a visualização de IO do grupo de partição atualizada; e

o novo nó de OSD primário de partição X é adaptado para: receber a solicitação de IO, e depois de executar a solicitação de IO, gerar uma solicitação de replicação, e enviar a nova solicitação de replicação para outro nó de OSD secundário na visualização de partição da partição atualizada, sendo que as etapas de gerar a solicitação de replicação e enviar a solicitação de replicação são as mesmas que as etapas anteriores 714 e 716.

[0131] Quando o nó de OSD secundário de Partição X se torna defeituoso depois de a solicitação de IO chegar ao nó de OSD primário de partição X, o procedimento de processamento de IO inclui o seguinte:

depois de o nó de OSD secundário de Partição X se tornar defeituoso,

o módulo de MDC é adicionalmente adaptado para: ao detectar, no processo de processamento da solicitação de IO, que o nó de OSD secundário de Partição X se torna defeituoso, marcar uma situação de partição do nó de OSD secundário de Partição X na visualização de partição como inconsistente, e enviar a visualização de partição atualizada de grupo de partição para o nó de OSD primário de partição X; e

o nó de OSD primário é adaptado para: depois de receber a visualização de partição atualizada do grupo de partição, enviar a solicitação de replicação para outro nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada, e ignorar o envio da solicitação de replicação para o nó de OSD secundário de Partição X cuja situação de partição é inconsistente.

[0132] No protocolo de comprometimento de duas fases, se um coordenador se torna defeituoso, o processamento de IO é interrompido, e continua apenas quando o coordenador volta ao normal. Nessa modalidade, depois de um nó de OSD primário se tornar defeituoso, o nó de MDC pode rapidamente determinar um novo nó de OSD primário por meio de votação, e rapidamente retomar o processamento de IO e, portanto, alta disponibilidade e forte tolerância à falha são alcançados.

[0133] Adicionalmente, no protocolo de comprometimento de duas fases, se um participante se torna defeituoso ou continuamente não tem resposta, outra solicitação de IO é continuamente bloqueada e finalmente falha devido ao esgotamento de tempo, e uma reversão precisa ser realizada. Nessa modalidade, se um nó secundário se torna defeituoso, o MDC instrui um nó primário a fazer alteração de visualização, de modo a isolar ou ignorar o nó de OSD defeituoso, e continua o processamento de solicitação de IO sem bloquear o processamento na outra solicitação de IO, que tem melhor tolerância à falha, e pode rapidamente manejar uma falha de nó e realizar um failback de nó. Por exemplo, uma falha de N réplicas em N+1 réplicas pode ser tolerada, o que aprimora adicionalmente o desempenho e a disponibilidade de um sistema de armazenamento. Um sistema com baixa disponibilidade inevitavelmente tem baixa extensibilidade, e devido ao fato de uma falha de nó de armazenamento ser comum em um sistema de armazenamento distribuído em larga escala, interação de protocolo complexa e massiva pode reduzir adicionalmente a

extensibilidade do sistema.

[0134] Adicionalmente, uma faixa de influência da falha de nó de armazenamento pode ser altamente diminuída por meio do controle de uma visualização de agrupamento em uma granularidade de partição, de modo que o sistema de armazenamento possa ser estendido em larga escala, e a extensibilidade de sistema seja aprimorada.

[0135] Depois de a solicitação de IO chegar no nó de OSD primário de partição X, o MDC descobre que um novo nó de OSD une o agrupamento e serve como um nó de OSD secundário da partição X, e o procedimento de processamento de IO inclui o seguinte:

o módulo de MDC notifica, quando determinar, no processo de processamento da solicitação de IO, que o novo nó de OSD une um agrupamento, o nó de OSD primário de partição X que o novo nó de OSD serve como um novo nó de OSD secundário no qual a partição X está localizado; depois da sincronização de dados de partição dados ser completada, atualiza uma visualização de partição e uma visualização de IO do grupo de partição; e instrui o nó de OSD primário de partição X para atualizar uma visualização de partição armazenada localmente no nó de OSD primário de partição X; e

o nó de OSD primário de partição X sincroniza dados de uma partição primária no nó de OSD primário de partição X para o novo nó de OSD secundário, e envia a solicitação de replicação para o novo nó de OSD secundário de acordo com a partição armazenada localmente.

[0136] Procedimento de processamento de falha de OSD:

[0137] A falha de nó de armazenamento é comum em um sistema de armazenamento em larga escala. Quando alguns nós de OSD são defeituosos, o sistema precisa ter capacidade para fornecer normalmente um serviço de IO. Nessa modalidade da presente invenção, o processamento em todas as solicitações depende de uma visualização de agrupamento mantida por um módulo de MDC e, quando um nó de OSD em um agrupamento se torna defeituoso, a visualização de agrupamento precisa também ser atualizada consequentemente, de modo que uma solicitação de IO possa ser processada efetivamente e de modo apropriado.

[0138] Para impedir que a falha de nó de OSD afete o processamento adequado na replicação da solicitação de IO, o seguinte

processamento precisa, em geral, ser realizado: Primeiro, o módulo de MDC detecta uma situação de um nó de OSD, e quando um nó de OSD se torna defeituoso, o módulo de MDC pode descobrir a falha em tempo; depois então de descobrir a falha de OSD, o módulo de MDC precisa realizar processamento em tempo para fazer uma alteração correta na visualização, e notificar um módulo de roteamento de IO e nó de OSD relacionados da alteração; então, o módulo de roteamento de IO e nó de OSD relacionados processam uma solicitação de IO correspondente de acordo com a visualização atualizada depois de receber uma notificação de atualização do MDC, de modo que o módulo e o nó possam obter a visualização atualizada em tempo, o que assegura que a solicitação de IO seja processada suavemente e efetivamente.

[0139] O módulo de MDC pode detectar a falha do nó de OSD conforme os seguintes dois modos: (1) o módulo de MDC é responsável pela detecção de falha em todos os nós de OSD, e cada nó de OSD envia regularmente uma mensagem de batimento cardíaco para o módulo de MDC, sendo que se um nó de OSD não enviar uma mensagem de batimento cardíaco para o módulo de MDC dentro de um período de tempo especificado, o módulo de MDC determina que o nó de OSD se torna defeituoso; e (2) nós de OSD enviam regularmente uma mensagem de batimento cardíaco entre si para detectar uma falha, sendo que se uma parte de detecção não recebe uma mensagem de batimento cardíaco de uma parte detectada dentro de um período de tempo especificado, a parte de detecção relata para o módulo de MDC que o nó de OSD correspondente se torna defeituoso.

[0140] Conforme descrito acima, tanto uma visualização de partição quanto uma visualização de IO são descritas em termos de um grupo de partição, e em geral, há múltiplas partições em um nó de OSD. Uma falha de um nó de OSD se relaciona à atualização de visualização para múltiplos grupos de partição que incluem as múltiplas partições, e a atualização de visualização para todos os grupos de partição são mutuamente independentes:

(1) quando uma partição no nó de OSD defeituoso inclui uma partição secundária, uma situação de partição do nó de OSD defeituoso em uma visualização de partição de um grupo de partição que inclui a partição secundária é marcada como inconsistente; ao mesmo tempo, um nó de OSD primário do grupo de partição é notificado de uma visualização de partição atualizada, e o nó

de OSD primário da partição processa, de acordo com a visualização de partição atualizada, a replicação de dados que correspondem a uma solicitação de IO; e

(2) quando uma partição no nó de OSD defeituoso inclui uma partição primária, o nó de OSD defeituoso que serve como um nó de OSD primário em uma visualização de partição de um grupo de partição que inclui a partição primária é definido como um novo nó de OSD secundário, uma situação de partição que corresponde ao novo nó de OSD secundário é marcada como inconsistente, um nó de OSD secundário cuja situação de partição é consistente é selecionado a partir de um nó de OSD secundário original na visualização de partição do grupo de partição que inclui a partição primária, e o nó de OSD secundário selecionai é definido como um novo nó de OSD primário; e então, o novo nó de OSD primário é notificado da atualização de visualização de partição, e outro nó de OSD secundário é notificado da atualização de visualização de IO, sendo que se o OSD primário do grupo de partição se torna defeituoso e as situações de partição de nós de OSD nas quais todas as partições secundárias estão localizadas são inconsistentes, nenhuma alteração é feita na visualização de partição e em uma visualização de IO; exige-se assegurar que uma réplica de partição primária tenha últimos dados completos, assegurando-se assim a consistência da replicação de dados.

[0141] O módulo de roteamento de IO relacionado e o nó de OSD processam uma solicitação de IO correspondente de acordo com a visualização atualizada depois de receber uma notificação de atualização do MDC pode incluir especificamente:

o novo nó de OSD primário replica, de acordo com a visualização de partição localmente armazenada atualizada, os dados que correspondem à solicitação de IO do módulo de roteamento de IO em um nó de OSD secundário no qual uma partição cuja situação de partição é consistente está localizada na visualização de partição localmente armazenada atualizada, ou em um nó de OSD secundário no qual uma partição cuja situação de partição é inconsistente está localizada na visualização de partição localmente armazenada atualizada, mas que está recuperando dados, de modo a isolar a falha, e assegurar solicitação de processamento de IO apropriado e ininterrupto, o que aprimora a tolerância à falha do sistema, e consequentemente aprimora o desempenho e a disponibilidade do sistema. Adicionalmente, uma faixa de influência do nó de

OSD falha pode ser diminuída por meio de controle de uma visualização de agrupamento em uma granularidade de partição, de modo que o sistema possa ser estendido em larga escala, e a extensibilidade de sistema é aprimorada.

[0142] Finalmente, depois de o nó de OSD defeituoso completar um failback e recuperação de dados, o módulo de MDC atualiza adicionalmente a visualização de partição e a visualização de IO, notifica um nó de OSD primário no qual uma partição na visualização de partição atualizada adicional está localizada da visualização de partição atualizada adicional, e envia a visualização atualizada de IO adicional para um nó de OSD secundário no qual uma partição na visualização de partição atualizada adicional está localizada, de modo que o módulo ou o nó de OSD que recebem a visualização de partição atualizada ou visualização de IO adicional atualize uma visualização de partição ou visualização de IO localmente armazenada, e processa, de acordo com a visualização de partição atualizada ou visualização de IO adicional, replicação dos dados que correspondem à solicitação de IO.

[0143] A visualização é atualizada em tempo, de modo que o nó de failback possa rapidamente unir o agrupamento para processar a solicitação de IO, o que aprimora o desempenho e eficiência do sistema.

[0144] Para facilitar o entendimento, uma modalidade específica é usada para descrição. A Figura 8A e Figura 8B mostram uma modalidade de um nó de procedimento de processamento de falha de OSD de acordo com a presente invenção. Essa modalidade é executada pelo módulo de MDC, o módulo de roteamento de IO, e o nó de OSD mencionado nas modalidades descritas na Figura 2A a Figura 2C e na Figura 5. Para facilitar a compreensão, um nó de OSDx, um nó de OSDy, e um nó de OSDz dentre os nós de OSD nessa modalidade são usados como um exemplo de descrição, sendo que o nó de OSDx, o nó de OSDy, ou o nó de OSDz podem ser qualquer nó de OSD de múltiplos nós de OSD em uma camada de protocolo de replicação distribuída nessa modalidade da presente invenção. Adicionalmente, para facilitar o entendimento, nessa modalidade, considera-se que o nó de OSDx seja um nó de OSD primário de um grupo de partição 1 (P1 em abreviação) e um nó de OSD secundário de um grupo de partição n (Pn em abreviação); sendo que o nó de OSDy é um nó de OSD primário da Pn e um nó de OSD secundário da P1; e o nó de OSDz é um nó de OSD secundário da Pn e um nó de OSD secundário da

P1. O nó de procedimento de processamento de falha de OSD nessa modalidade inclui especificamente as seguintes etapas:

[0145] 802/804/806 O nó de OSDx, o nó de OSDy, e o nó de OSDz enviam separadamente e regularmente uma mensagem de batimento cardíaco para o módulo de MDC primário.

[0146] 808. O nó de OSDx se torna defeituoso.

[0147] Em uma aplicação real, podem ser várias falhas de software, hardware ou rede. Por exemplo, devido ao fato de uma reinicialização de processo de programa causado por um BUG de software, uma desconexão de rede temporária, uma reinicialização de servidor, e similares, o nó de OSD não pode processar a solicitação de IO, e não pode implantar funções de armazenamento e replicação de dados.

[0148] 810. Se o módulo de MDC detecta que o nó de OSDx não envia a mensagem de batimento cardíaco dentre de um tempo predeterminado, o módulo de MDC determina que o nó de OSDx se torna defeituoso.

[0149] Em uma aplicação real, conforme descrito acima, o MDC pode também determinar, de outra maneira, que o nó de OSDx se torna defeituoso. Por exemplo, um nó de OSD funcional notifica o módulo de MDC que um nó de OSD se torna defeituoso.

[0150] 812. O módulo de MDC realiza a atualização de visualização de acordo com uma situação de falha.

[0151] Em uma maneira de implantação específica, o MDC atualiza uma visualização de agrupamento de um grupo de partição correspondente de acordo com um grupo de partição que inclui uma partição no nó de OSD defeituoso determinado. Nessa modalidade, partições que incluem uma partição no nó de OSDx defeituoso incluem a P1 e a Pn e, portanto, o MDC precisa atualizar as visualizações de agrupamento da P1 e da Pn, que podem incluir especificamente:

(1) Atualizar uma visualização de OSD: Uma situação do nó de OSDx é atualizada de um estado "DENTRO (IN)" e "ASCENDENTE (UP)" para um estado "DENTRO (IN)" e "DESCENDENTE (DOWN)".

(2) Alterar uma visualização de partição: Para a P1, o nó de OSDy é definido como um nó de OSD primário da P1 (o primeiro nó de OSD secundário cuja situação de partição é consistente é selecionado a partir de uma lista de nó

de OSD secundário da P1, e é definido como um nó de OSD primário do grupo de partição), o OSDx é definido como um nó de OSD secundário da P1, e uma situação de partição correspondente do mesmo é atualizada para "inconsistente"; para a Pn, uma situação de partição do OSDx secundário da Pn é alterada para inconsistente.

(3) Atualizar uma visualização de IO: Para a P1, o nó de OSDx original é substituído pelo nó de OSDy como um nó de OSD primário em uma visualização de IO da P1; para a Pn, devido ao fato de o nó de OSDx defeituoso servir meramente como um nó de OSD secundário da Pn, e o nó de OSDy primário da Pn não se tornar defeituoso, uma visualização de IO da Pn não é atualizada.

[0152] 814. Notificar o nó de OSDy (que serve como um nó de OSD primário da P1 depois de atualizar e ainda serve como um nó de OSD primário da Pn) da atualização de visualizações de partição da P1 e da Pn, sendo que a atualização inclui: definir o nó de OSDy como um nó de OSD primário da P1, definir o OSDx como um nó de OSD secundário da P1 e uma situação de partição correspondente do OSDx para "inconsistente", e mudar uma situação de partição do nó de OSDx secundário da Pn para "inconsistente".

[0153] 816. Notificar o nó de OSDz da atualização de uma visualização de IO da P1, isto é, notificar o nó de OSDz que o nó de OSD primário da P1 é substituído pelo nó de OSDy.

[0154] 818. Notificar o módulo de roteamento de IO da atualização da visualização de IO da P1, isto é, notificar o módulo de roteamento de IO que o nó de OSD primário da P1 é substituído pelo nó de OSDy.

[0155] 820. O nó de OSDy processa uma notificação do módulo de MDC, atualiza informações de visualização localmente armazenadas (uma visualização de partição e uma visualização de IO), e processa, de acordo com uma última visualização notificada pelo módulo de MDC, a replicação dos dados que correspondem à solicitação de IO.

[0156] Em uma maneira de implantação específica, como o nó de OSD primário da P1 depois de atualização, o nó de OSDy atualiza uma visualização de partição e uma visualização de IO da P1; como o nó de OSD original da Pn, o nó de OSDy atualiza uma visualização de partição da Pn.

[0157] Para uma operação de IO na P1, depois de receber a

solicitação de IO encaminhada pelo módulo de roteamento de IO, o nó de OSDy executa a solicitação de IO, gera uma solicitação de replicação, e envia a solicitação de replicação para o nó de OSD secundário da P1 na visualização de partição atualizada, isto é, o nó de OSDz, sendo que uma situação de partição correspondente do nó de OSDz é "consistente". Divisão de o nó de OSDx servir como um novo nó de OSD secundário da P1, e uma situação de partição do nó de OSDx ser "inconsistente", o nó de OSDy não envia mais a solicitação de replicação para o nó de OSDx, que implanta isolamento de falha, e não afeta a solicitação de processamento de IO contínua na P1.

[0158] Para uma operação de IO na Pn, depois de receber a solicitação de IO encaminhada pelo módulo de roteamento de IO, o nó de OSDy executa a solicitação de IO, gera uma solicitação de replicação, e envia a solicitação de replicação para o nó de OSD secundário da Pn na visualização de partição atualizada, isto é, o nó de OSDz, sendo que uma situação de partição correspondente do nó de OSDz é "consistente". Devido ao fato de o nó de OSDx servir como um novo nó de OSD secundário da Pn, e uma situação de partição do nó de OSDx ser "inconsistente", o nó de OSDy não envia mais a solicitação de replicação para o nó de OSDx, que implanta isolamento de falha, e não afeta a solicitação de processamento de IO contínua na P1.

[0159] No protocolo de comprometimento de duas fases, se um participante se torna defeituoso ou continuamente não tem resposta, outra solicitação de IO é continuamente bloqueada e finalmente falha devido ao esgotamento de tempo, e uma reversão precisa ser realizada. Nessa modalidade, se um nó de OSD secundário se torna defeituoso, o MDC instrui um nó primário a realizar uma alteração de visualização, de modo a ignorar o nó defeituoso, isolar o nó defeituoso e continuar a solicitação de processamento de IO sem bloquear o processamento na outra solicitação de IO, que tem melhor tolerância à falha e disponibilidade.

[0160] 822. Processar uma notificação de visualização do módulo de MDC, e atualizar informações de visualização de IO armazenadas localmente.

[0161] 824. Processar a notificação de visualização do módulo de MDC, atualizar as informações de visualização de IO armazenadas localmente, e realizar roteamento e encaminhamento de solicitação de IO de acordo com uma última visualização de IO notificada pelo módulo de MDC.

[0162] Em um modo de implantação específico, para processamento de IO no P1, se o nó de OSDx primário original no qual o P1 é localizado se tornar defeituoso em um processo de processamento de IO do P1, o MDC atualiza a visualização de partição do P1 em tempo, e o módulo de roteamento de IO encaminha, de acordo com o nó de OSDy primário recém-selecionado do P1 na visualização de partição atualizada, a solicitação de IO para o nó de OSDy recém-selecionado novamente.

[0163] Nessa modalidade da presente invenção, após um nó de OSD primário se tornar defeituoso, o nó de MDC pode determinar rapidamente um novo nó primário por meio de votação, e retomar rapidamente o processamento de IO, e se um nó secundário se tornar defeituoso, o MDC instrui o nó primário a realizar uma alteração de visualização, de modo a isolar ou ignorar o nó de OSD defeituoso, e continuar o processamento de solicitação de IO sem bloquear o processamento em outra solicitação de IO, que tem melhor tolerância contra falha, e pode rapidamente lidar com uma falha de nó. Por exemplo, uma falha de N réplicas em N+1 réplicas pode ser tolerada, que aperfeiçoa adicionalmente desempenho e disponibilidade de um sistema de armazenamento. Um sistema com baixa disponibilidade tem inevitavelmente baixa extensibilidade, e visto que uma falha de nó de armazenamento é comum em um sistema de armazenamento distribuído em larga escala, interação de protocolo complexa e massiva pode reduzir adicionalmente extensibilidade de sistema. Além disso, uma faixa de influência da falha de nó de armazenamento pode ser bastante reduzida por meio de controle de uma visualização de agrupamento em uma granularidade de partição, de modo que o sistema de armazenamento possa ser estendido em uma larga escala, e a extensibilidade de sistema seja aperfeiçoada.

[0164] Procedimento de failback de nó de OSD:

[0165] Uma nova operação de modificação de dados pode ocorrer durante uma falha de um nó de OSD e, portanto, antes de o nó de OSD defeituoso retornar ao normal e se juntar novamente a um agrupamento para fornecer um serviço, a recuperação de dados e sincronização precisam ser primeiramente realizados, de modo a fazer com que o nó de OSD defeituoso retorne para um estado consistente com aquele de uma réplica primária.

[0166] Em uma modalidade da presente invenção, o procedimento

de failback de nó de OSD pode ser dividido em três fases:

(1) um nó de OSD secundário sincroniza, com um nó de OSD primário, modificação de dados realizada pelo nó de OSD primário durante a falha, que é um processo de sincronização incrementar e, certamente, em uma aplicação real, todos os dados de uma partição podem ser sincronizados de acordo com uma situação real.

(2) após o nó de OSD secundário retornar para um estado consistente com aquele do nó de OSD primário, um módulo de MDC altera uma visualização de agrupamento.

(3) O módulo de MDC notifica cada módulo e nó da visualização de agrupamento atualizada, de modo que cada módulo e nó processe encaminhamento ou replicação de solicitação de IO de acordo com a visualização de agrupamento atualizada notificada.

[0167] Em uma aplicação real, se o nó de OSD primário enviar uma solicitação de replicação ao nó de failback após receber uma solicitação de IO em um processo de recuperação de dados, o procedimento de failback de nó de OSD pode incluir adicionalmente a fase a seguir:

[0168] O nó de OSD secundário reproduz de volta um registro que é gravado após a solicitação de replicação do nó de OSD primário ser recebida e em um processo de sincronização de dados com o OSD primário, e escreve dados que estão no registro. Desse modo, pode-se garantir que, em um processo de failback, todos os dados do nó de OSD de failback sejam consistentes com aqueles do nó de OSD primário, aperfeiçoando adicionalmente, desse modo, a consistência de dados entre o nó de OSD primário e o nó de OSD secundário.

[0169] Para um procedimento específica de um processo de sincronização de dados, referir a uma modalidade específica fornecida na Figura 10A e Figura 10B a seguir.

[0170] Um processo de notificação de atualização de visualização de agrupamento pode incluir os casos a seguir:

(1) Se o nó de OSD é um nó de OSD secundário de um grupo de partição antes da falha, apenas uma visualização de partição pode ser alterada, em que uma situação de partição do nó de OSD secundário que corresponde à visualização de partição é mudada para um estado "consistente", e a

visualização de partição alterada é notificada ao nó de OSD primário.

(2) Se o nó de OSD for um OSD primário de um grupo de partição antes da falha, tanto uma visualização de partição como uma visualização de IO são alteradas, o nó de OSD anterior é definido como um novo nó de OSD primário do grupo de partição, o OSD primário atual é definido como um nó de OSD secundário do grupo de partição, o novo nó de OSD primário é instruído para alterar a visualização de partição, e um módulo de roteamento de IO e o nó de OSD secundário são instruídos para alterar a visualização de IO.

[0171] Em uma aplicação real, se o nó de OSD defeituoso não retornar ao normal dentro de um período de tempo especificado, o módulo de MDC expulsa o nó de OSD do agrupamento, e migra uma partição distribuída no nó de OSD para outro nó de OSD. Para um procedimento específico, referir a uma modalidade específica em que um nó de OSD sai de um agrupamento fornecido na Figura 11A e Figura 11B a seguir.

[0172] Para facilitar o entendimento, a seção a seguir usa uma modalidade específica para descrição. A Figura 9A, a Figura 9B e a Figura 9C mostram uma modalidade de um procedimento de processamento de failback de nó de OSD de acordo com a presente invenção. Essa modalidade é executada pelo módulo de MDC, pelo módulo de roteamento de IO e pelo nó de OSD mencionado nas modalidades descritas na Figura 2A a Figura 2C e Figura 5. Para facilitar o entendimento, é previsto nessa modalidade que um exemplo no qual um grupo de partição que inclui uma partição no nó de OSD de failback inclui um grupo de partição 1 (P1 para abreviar) e um grupo de partição n (Pn para abreviar) seja usado para descrição, e o procedimento de processamento de falha de nó de OSD nessa modalidade inclui especificamente as etapas a seguir:

[0173] 902. O nó de OSD de failback solicita o módulo de MDC para informações de visualização de agrupamento no nó de OSD de failback, em que a solicitação porta um ID de OSD do nó de OSD.

[0174] Para um modo de implantação específico, referir ao processo anterior descrito na etapa 608 na Figura 6.

[0175] 904. O módulo de MDC consulta uma visualização de partição de acordo com o ID de OSD para adquirir informações de partição sobre o OSD.

[0176] Em um modo de implantação específico, o MDC consulta separadamente, de acordo com o ID de OSD do OSD de failback, visualizações de partição correspondentes ao P1 e ao Pn no OSD de failback, para obter separadamente informações de partição do P1 einformações de partição do Pn. As informações de partição podem incluir uma visualização de IO, e podem incluir adicionalmente uma situação de partição.

[0177] 906. Retornar as informações de partição para o nó de OSD.

[0178] 908/910. Começar um processo de recuperação de dados para um nó de OSD primário do grupo de partição de acordo com as informações de partição retornadas pelo MDC, e adquirir separadamente informações de entrada ausentes durante uma falha de nós de OSD primários do P1 e do Pn, em que as informações de entrada portam uma Seq ID.

[0179] Em uma modalidade específica da presente invenção, se um nó de OSD primário (por exemplo, o nó de OSD primário do Pn) receber uma solicitação de escrita de IO no processo de failback, o nó de OSD primário precisa enviar uma solicitação de replicação a todos os nós de OSD secundários do Pn, e o procedimento de failback pode incluir adicionalmente as etapas a seguir 912 a 916 e a etapa 918.

[0180] 912. O nó de OSD primário do Pn recebe uma solicitação de escrita de IO de um hospedeiro durante um failback.

[0181] 914. Enviar informações de IO de replicação a um nó de OSD secundário no qual uma partição é localizada.

[0182] Em um modo de implantação específico, para as etapas anteriores 912 a 914, referir ao procedimento de operação de IO anterior na Figura 7A e Figura 7B.

[0183] 916. Escrever informações de chave de IO de replicação em um registro (log).

[0184] Em um modo de implantação específico, dados que correspondem à solicitação de IO também podem ser escritos no registro, e para as informações de chave de IO nessa etapa, referir à descrição do procedimento de operação de IO anterior na Figura 7A e Figura 7B.

[0185] 918. Enviar repetidamente uma mensagem de solicitação de dados de acordo com informações de quantidade de entrada de diferença retornadas pelo nó de OSD primário e de acordo com uma entrada, de modo a

garantir que o processo de recuperação de dados seja concluído.

[0186] Para um modo de implantação específico, referir a um procedimento de recuperação de dados na Figura 11A e Figura 11B a seguir.

[0187] 920. Escrever IO de acordo com as informações de IO gravadas no registro e de acordo com a gravação do registro.

[0188] Em um modo de implantação específico, após concluir recuperação de dados e sincronização com o nó de OSD primário, o nó de OSD de failback escreve, para recursos de armazenamento físico gerenciados pelo nó de OSD de failback e de acordo com as informações de IO gravadas no registro, dados que correspondem à solicitação de replicação do OSD primário do Pn.

[0189] A solicitação de IO é escrita no registro no processo de failback e, então, uma solicitação de IO gerada durante a falha é escrita após dados ausentes durante a falha serem recuperados, que pode garantir que as sequências em que o nó de OSD primário e o nó de OSD secundário executam todas as operações de IO sejam consistentes, e aperfeiçoam adicionalmente consistência de cópia de segurança de dados.

[0190] 922/924. Concluir recuperação de dados, e notificar separadamente os nós de OSD primários do P1 e do Pn.

[0191] Em um modo de implantação específico, o módulo de MDC pode ser instruído, em dois modos, para executar atualização de uma visualização de agrupamento, em que no Modo 1, após a recuperação de dados ser concluída, o nó de OSD de failback notifica os nós de OSD primários do P1 e do Pn, de modo que o nó de OSD primário instrua o nó de MDC para atualizar a visualização de agrupamento. Para o Modo 2, referir à etapa a seguir 930.

[0192] Em uma aplicação real, antes de instruir para atualizar a visualização de agrupamento, o nó de OSD de failback pode determinar adicionalmente uma situação de partição da partição no nó de OSD de failback, e aciona um procedimento de atualização de visualização de agrupamento após determinar que a situação de partição é inconsistente. O nó de OSD de failback pode adquirir adicionalmente informações de situação de partição reunindo-se as informações de partição retornadas na etapa anterior 906.

[0193] 926/928. Os nós de OSD primários do P1 e do Pn instruem separadamente o MDC a alterar a visualização de partição, em que uma

notificação porta um ID de grupo de partição, um ID de um OSD secundário e uma versão de vista.

[0194] Em um modo de implantação específico, a notificação é enviada ao módulo de MDC para exigir que o módulo de MDC atualize a situação de partição da partição no nó de OSD de failback para consistente, em que a notificação porta o ID de grupo de partição do grupo de partição que inclui a partição no nó de OSD, o ID do nó de OSD secundário (ou seja, o nó de OSD de failback), e a versão de vista. O ID de grupo de partição, no presente documento, é usada para marcar uma visualização a ser atualizada do grupo de partição, o ID de OSD é usada para marcar o nó de OSD defeituoso, e a versão de vista no presente documento é uma versão de vista de uma última visualização de partição localmente armazenada nos nós de OSD primários em que o P1 e o Pn são localizados, em que uma função da versão de vista é que o MDC realiza processamento de atualização de visualização de agrupamento após receber a notificação e determinar que a versão de vista da visualização de partição na notificação é consistente com uma versão de vista de uma última visualização de partição localmente mantida pelo MDC, garantindo, desse modo, que visualizações de agrupamento vistas por todos os módulos ou nós no procedimento de processamento de IO sejam consistentes, e aperfeiçoando consistência de backup de dados.

[0195] Em uma aplicação real, se for determinado que a versão de vista da visualização de partição na notificação é inconsistente com a versão de vista da última visualização de partição localmente mantida pelo MDC, a última visualização de partição é enviada adicionalmente aos nós de OSD primários nos quais o P1 e o Pn são localizados, e após ser determinado que os dados primários e os dados secundários do P1 e do Pn são consistentes, a visualização de agrupamento é atualizada.

[0196] 928. Instruir o MDC para alterar a visualização de partição, e atualizar uma situação do nó de OSD de failback para consistente, em que uma instrução porta o ID de grupo de partição, o ID do OSD secundário e a versão de vista.

[0197] 930. Concluir a recuperação de dados, e instruir o MDC primário a alterar a visualização de partição, em que uma notificação porta o ID de grupo de partição, o ID do OSD secundário e a versão de vista.

[0198] Em um modo de implantação específico, após concluir a recuperação de dados, o nó de OSD de failback envia a notificação ao módulo de MDC, para instruir o módulo de MDC a atualizar a situação de partição da partição no nó de OSD de failback para consistente. Uma diferença entre o Modo 1 e o Modo 2 anteriores está em que a versão de vista no presente documento é uma versão de vista de partição de uma última visualização de partição ou uma versão de vista de IO de uma última visualização de IO armazenada localmente no nó de OSD de failback (se a versão de vista é a visualização de partição ou a visualização de IO depende especificamente na possibilidade de que o nó de OSD de failback seja um nó de OSD primário ou um nó secundário antes da falha), e após determinar, de acordo com a versão de vista na notificação, que a versão de vista correspondente mantida localmente pelo módulo de MDC é consistente, o módulo de MDC realiza processamento de atualização de visualização de agrupamento.

[0199] 932. Atualizar uma situação de partição do nó de OSD de failback na visualização de partição para consistente.

[0200] Em um modo de implantação específico, as situações de partição que correspondem ao nó de OSD de failback nas visualizações de partição do P1 e do Pn no nó de OSD de failback são atualizadas para "consistente".

[0201] 934. Determinar, por meio de comparação de visualização, se o nó de OSD de failback é um nó de OSD primário de algumas partições antes da falha.

[0202] Em um modo de implantação específico, é determinado, comparando-se as últimas visualizações de partição que são do P1 e do Pn no nó de OSD de failback e que são mantidas localmente pelo módulo de MDC com visualizações de partição inicializadas do P1 e do Pn, se o nó de OSD de failback é o nó de OSD primário do P1; ou é determinado, comparando-se as últimas visualizações de IO que são do P1 e do Pn no nó de OSD de failback e que são mantidas localmente pelo módulo de MDC com visualizações de IO inicializadas do P1 e do Pn, se o nó de OSD de failback é o nó de OSD primário do P1.

[0203] 936. Redefinir o nó de failback como um nó de OSD primário, e atualizar a visualização de partição.

[0204] Em um modo de implantação específico, para o P1, o MDC

redefine o nó de OSD de failback como um novo nó de OSD primário do P1, define o nó de OSD primário atual no qual o P1 é localizado como um novo nó de OSD secundário do P1, e atualiza a visualização de partição do P1; para o Pn, visto que o nó de OSD de failback é originalmente um nó de OSD secundário do Pn, para o Pn, um problema de mudança de identidade de primário/secundário do nó de OSD de failback não está envolvido.

[0205] 938. Se o nó de OSD de failback for originalmente um nó primário, enviar uma última Visualização de Partição ao nó de OSD de failback.

[0206] Em um modo de implantação específico, visto que o nó de OSD de failback é originalmente um nó primário do P1, e o nó de OSD de failback é definido como um novo nó de OSD primário do nó de OSD de failback de acordo com a etapa 937, a última Visualização de Partição atualizada do P1 precisa ser enviada para o novo nó de OSD primário do nó de OSD de failback, ou seja, o nó de OSD de failback. Visto que o nó de OSD de failback não é originalmente um nó primário do Pn, uma última visualização de partição do Pn pode não ser enviada ao nó de OSD de failback.

[0207] 940/942. Instruir o OSD secundário para atualizar uma visualização de IO local.

[0208] Em um modo de implantação específico, de acordo com as últimas visualizações de partição ou visualizações de IO do P1 e do Pn, novos nós primários de OSD do P1 e do Pn são adquiridos separadamente, as últimas visualizações de IO do P1 e do Pn são respectivamente enviadas ao novo nó de OSD primário do P1 e o nó de OSD primário do Pn (visto que para o Pn, uma mudança do nó de OSD primário não está envolvida, a visualização de IO ainda é enviada ao nó de OSD primário original do Pn).

[0209] 944. O nó de OSD de failback determina se OSDs primário na visualização de partição e a visualização de IO localmente armazenada são consistentes, determina se o nó de OSD de failback é definido como um nó primário e atualiza a visualização de IO .

[0210] Após receber a última visualização de partição, o nó de OSD de failback determina se OSDs primários na última visualização de partição e a visualização de IO localmente armazenada são consistentes, determina se o nó de OSD de failback é definido como um nó primário, e se os OSDs primário na última visualização de partição e a visualização de IO localmente armazenada

forem consistentes, e o nó de OSD de failback é definido como um nó primário, atualiza a visualização de IO e a visualização de partição localmente armazenada, e processa, de acordo com a visualização de partição atualizada e a visualização de IO atualizada, a replicação de dados relacionados à solicitação de IO.

[0211] 946. Atualizar a visualização de IO, e se o OSD primário for definido como um OSD secundário, excluir a visualização de partição.

[0212] Em um modo de implantação específico, nós de OSD primários atuais do P1 e do Pn (ou seja, nós de OSD que existem após a recuperação de dados ser concluída e antes de a atualização iniciar) recebem a visualização de IO atualizada, e atualizam as respectivas visualizações de IO localmente armazenadas. Para o nó de OSD primário atual do P1, visto que o nó de OSD primário atual foi definido pelo módulo de MDC como um novo nó de OSD secundário do P1, o nó de OSD primário atual do P1 exclui a visualização de partição localmente armazenada, e processa a replicação da solicitação de IO de acordo com a visualização de IO localmente armazenada atualizada.

[0213] Nessa modalidade da presente invenção, após um nó de OSD se tornar defeituoso, um MDC instrui outro nó relacionado a realizar atualização de visualização, de modo a isolar ou ignorar o nó de OSD defeituoso, continuar processamento de solicitação de IO sem bloquear processamento em outra solicitação de IO, atualizar uma visualização após um failback de nó e notificar cada nó relacionado, de modo que o nó de failback possa se unir de novo, de maneira rápida a um agrupamento novamente para trabalho, que pode processar rapidamente uma falha de nó e renegociação, tem melhor tolerância contra falha, e aperfeiçoa o desempenho e disponibilidade de um sistema de armazenamento. Um sistema com baixa disponibilidade tem inevitavelmente baixa extensibilidade, e visto que uma falha de nó de armazenamento é comum em um sistema de armazenamento de distribuído de larga escala, interação de protocolo complexa e massiva pode reduzir adicionalmente extensibilidade de sistema. Além disso, uma faixa de influência da falha de nó de armazenamento pode ser bastante reduzido por meio de controle de uma visualização de agrupamento em uma granularidade de partição, de modo que o sistema de armazenamento possa ser estendido em uma grande escala, e a extensibilidade de sistema é aperfeiçoada.

[0214] Procedimento de recuperação de dados:

[0215] A seguir há uma descrição de um procedimento de processamento de recuperação de dados no processo de processamento de failback de OSD anterior na Figura 9A, Figura 9B, e Figura 9C usando-se uma modalidade específica. A fim de facilitar o entendimento, é previsto nessa modalidade que um exemplo no qual um grupo de partição que inclui uma partição no nó de OSD de failback inclui uma partição 1 (P1 para abreviar) e uma partição n (Pn para abreviar) seja usado para a descrição. Conforme mostrado na Figura 10A e na Figura 10B, o procedimento de processamento de falha de nó de OSD, nessa modalidade, inclui especificamente as etapas a seguir:

[0216] 1002. O nó de OSD de failback adquire localmente uma maior Seq ID em uma entrada gravada de cada partição.

[0217] Conforme mostrado na etapa 712 e na etapa 722 na modalidade anterior na Figura 7A e na Figura 7B, um nó de OSD em um sistema grava uma entrada para cada operação de IO em uma partição em um processo de processamento de IO. Conforme descrito acima, a entrada inclui um tipo de operação de IO, um ID de partição, uma Seq ID e uma chave, e a entrada pode incluir adicionalmente informações de situação, que são usadas para descrever se a operação é bem-sucedida. Além disso, para uma operação de escrita de IO, a entrada pode incluir adicionalmente o desvio e comprimento anteriores. Por exemplo, nessa modalidade, maiores Seq IDs para operações de escrita de IO no P1 e no Pn são adquiridas separadamente.

[0218] 1004/1006. Solicitar separadamente nós de OSD primários do P1 e do Pn para adquirir uma entrada que o OSD carece, em que a solicitação porta um ID de grupo de partição e uma respectiva maior Seq ID.

[0219] Cenário 1: Uma maior Seq ID de um OSD secundário está dentro de uma faixa de uma entrada gravada por um OSD primário.

[0220] 1008/1010. Os nós de OSD primários do P1 e do Pn enviam separadamente entradas que o nó de OSD de failback carece para o nó de OSD de failback.

[0221] Em um modo de implantação específico, se uma maior Seq ID do nó de OSD de failback for 1,6, e uma maior Seq ID atual do OSD primário for 1,16, dez entradas correspondentes com Seq IDs de 1,7 a 1,16 são enviadas para o OSD secundário, que é meramente um exemplo fornecido para facilitar o

entendimento. Em uma aplicação real, modos ou regras de numeração de Seq ID podem ser diferentes, e os nós de OSD primários do P1 e do Pn podem carecer de diferentes entradas.

[0222] 1012/1014. O nó de OSD de failback realiza repetidamente sincronização de dados de acordo com a entrada adquirida na etapa anterior, em que a solicitação porta informações de chave de IO tais como uma chave, um desvio e um comprimento.

[0223] Em um modo de implantação específico, o nó de OSD de failback envia, de acordo com as entradas adquiridas, solicitações de sincronização de dados uma por uma em bateladas para os nós de OSD primários nos quais o P1 e o Pn são localizados, em que a solicitação porta as informações de chave de IO tais como a chave, o desvio e o comprimento.

[0224] 1016. Enviar dados correspondentes.

[0225] Em um modo de implantação específico, os nós de OSD primários do P1 e do Pn enviam, de acordo com informações na solicitação de sincronização de dados adquiridos, dados que correspondem a cada entrada para o nó de OSD de failback.

[0226] Cenário 2: Uma maior Seq ID de um OSD secundário não está dentro de uma faixa de uma entrada gravada por um OSD primário e é menos do que uma menor Seq ID do OSD primário.

[0227] 1018. Enviar uma menor Seq ID do OSD primário e nenhuma entrada.

[0228] Em um modo de implantação específico, em um caso em que o nó de OSD primário do Pn determina que a maior Seq ID do nó de OSD de failback não está dentro da faixa da entrada gravada pelo OSD primário, ou seja, a maior Seq ID é menor do que uma menor Seq ID do OSD primário do Pn, a menor Seq ID do OSD primário do Pn e nenhuma entrada são enviados para o nó de OSD de failback, que ajuda o nó de OSD de failback a determinar se o nó de OSD primário do Pn não escreve dados ou escreve dados excessivos, a tal ponto that a recuperação de dados não pode ser concluída por meio de sincronização incrementar.

[0229] 1020. Solicitar repetidamente sincronização de dados de partição, em que uma solicitação porta o ID de grupo de partição.

[0230] O nó de OSD de failback solicita ao nó de OSD primário do

Pn sincronização de dados de toda uma partição, em que, por exemplo, os dados de uma partição primária no OSD primário do Pn, nessa modalidade, são sincronizados com dados de uma partição secundária do Pn no nó de OSD de failback, e a solicitação porta o ID de grupo de partição. Em um modo de implantação específico, um volume de dados de uma partição é geralmente um tanto grande, de modo que os dados não possam ser totalmente transmitidos usando-se uma solicitação e, além disso, o nó primário não conhece uma capacidade de IO do nó de failback, mas envia continuamente os dados ao nó de OSD de failback, e o nó de failback pode fracassar em processar os dados; portanto, o nó primário envia os dados para o nó de failback apenas quando o nó de failback solicitar sincronização de dados. O nó de OSD de failback envia repetidamente uma solicitação de sincronização para o nó de OSD primário do Pn de acordo com uma situação, para sincronizar os dados em toda uma partição até que todos os dados na partição sejam sincronizados. Em uma aplicação real, toda a partição pode ser sincronizada de outra maneira, que não é limitada na presente invenção.

[0231] 1022. Enviar dados que correspondem a uma ou mais chaves.

[0232] O nó de OSD primário do Pn envia, de acordo com uma solicitação de sincronização enviada pelo nó de OSD de failback cada vez, os dados que correspondem à uma ou mais chaves.

[0233] O procedimento no qual um nó de OSD sai de um agrupamento após estar defeituoso:

[0234] Se um nó de OSD ainda não puder retornar ao normal e se unir novamente a um agrupamento após a falha durar mais do que um limite de tempo predefinido (por exemplo, 5 minutos) ou uma falha de hardware ocorrer no nó de OSD, o nó de OSD defeituoso precisa ser expulso do agrupamento, de modo a garantir confiabilidade de dados.

[0235] Sair do agrupamento pelo nó de OSD é um processo de redistribuição de partição e migração de dados, em que em redistribuição de partição, o equilíbrio de cada nó e segurança de réplica precisam ser considerados. O processamento de IO em um processo de migração de dados é consistente com o processamento em um procedimento de failback e um procedimento de recuperação de dados; após a migração de dados estar

concluída, uma réplica primária e uma réplica secundária alcançam um estado consistente, e um processo em que um MDC realiza atualização de visualização e notificação é consistente com o processamento de atualização de visualização realizado após um failback ser concluída. Um processo no qual cada nó de OSD relacionado realiza processamento de encaminhamento ou replicação em uma solicitação de IO de acordo com uma visualização atualizada é também consistente com o processamento de IO realizado por cada nó de OSD de acordo com uma última visualização após um failback estar concluída.

[0236] Para facilitar o entendimento, a seção a seguir usa uma modalidade específica para a descrição. A Figura 11A e Figura 11B mostram uma modalidade de um procedimento de processamento de failback de nó de OSD de acordo com a presente invenção. Essa modalidade é executada pelo módulo de MDC, pelo módulo de roteamento de IO e pelo nó de OSD mencionado nas modalidades descritas na Figura 2A a Figura 2C e Figura 5. Para facilitar entendimento, é previsto que nessa modalidade, nós de OSD relacionados sejam um nó de OSD1, um nó de OSD2, e um nó de OSDn, em que o nó de OSD1 é um nó de OSD primário de um grupo de partição 1 (P1 para abreviar) e um nó de OSD secundário de um grupo de partição 2 (P2 para abreviar), o nó de OSD2 é um nó de OSD secundário do P1, e o nó de OSDn é um nó de OSD primário do P2. O procedimento de processamento no qual um nó de OSD sai após estar defeituoso nessa modalidade inclui especificamente as etapas a seguir:

[0237] 1100. O nó de OSD1 se torna defeituoso.

[0238] 1102. Quando o módulo de MDC constata que a falha do OSD1 dura mais do que um limite predeterminado ou uma falha de hardware ocorre no OSD1, o MDC expulsa o OSD1 de um agrupamento, altera uma visualização, e migra partições (no presente documento, as partições são uma partição primária do P1 e uma partição secundária do P2) no nó de OSD1 para outros nós de OSD, por exemplo, o nó de OSD2 e o nó de OSDn nessa modalidade.

[0239] 1104. O módulo de MDC notifica o nó de OSD2 de atualização de visualização: o nó de OSD2 é definido como um nó primário do P1, e se torna um nó secundário do P2.

[0240] 1106. O módulo de MDC notifica o nó de OSDn de

atualização de visualização: o nó de OSDn é um nó primário do P2, e se torna um nó secundário do P1.

[0241] 1108. O nó de OSD2 solicita ao nó de OSDn para sincronizar os dados do P2.

[0242] Visto que o nó de OSDn é o nó de OSD primário do P2, o nó de OSD2 solicita ao nó de OSDn para sincronizar dados de uma partição primária do P2 no OSDn, de modo que os dados de uma partição secundária de P2 no OSD2 sejam consistentes com os dados da partição primária do P2 no OSDn. Um procedimento de sincronização específica é similar a um procedimento de sincronização de dados de toda uma partição no procedimento de recuperação de dados anterior mostrado na Figura 10A e na Figura 10B, e os detalhes não são descritos novamente no presente documento.

[0243] 1110. O nó de OSDn solicita o nó de OSD2 para sincronizar dados do P1.

[0244] Visto que o nó de OSD2 é o nó de OSD secundário original do P1, após o nó de OSD1 se tornar defeituoso, como um novo nó de OSD secundário do P1, o nó de OSDn pode sincronizar dados do P1 apenas com o OSD2, e o nó de OSDn solicita ao nó de OSD2 para sincronizar dados de uma partição primária do P1 no OSD2, de modo que os dados de uma partição secundária do P1 no OSDn sejam consistentes com os dados da partição primária do P1 no OSD2.

[0245] 1112. Concluir sincronização de dados de partição.

[0246] 1114. O nó de OSD2 notifica o módulo de MDC que a migração de dados do P2 foi concluída.

[0247] 1116. O nó de OSDn notifica o módulo de MDC que a migração de dados do P1 foi concluída.

[0248] 1118. O módulo de MDC realiza atualização de visualização de acordo com uma notificação correspondente.

[0249] Um princípio de atualização de visualização específico é o mesmo que aquele descrito nos procedimentos anteriores, um processo de atualização é o mesmo que aquele descrito nos procedimentos anteriores, e os detalhes não são descritos novamente no presente documento.

[0250] 1120. Notificar atualização de visualização: uma situação de partição do nó de OSDn secundário do P1 é consistente.

[0251] 1122. Notificar atualização de visualização: uma situação de partição do nó de OSD2 secundário do P2 é consistente.

[0252] 1124. O nó de OSD2 e o nó de OSDn processam, de acordo com uma última visualização, replicação de dados que correspondem a uma solicitação de IO.

[0253] O procedimento no qual um novo nó de OSD se une a um agrupamento:

[0254] Um novo nó se une a um agrupamento, e uma partição que é originalmente distribuída em outro nó de OSD precisa ser migrada para o novo nó de OSD que se une ao agrupamento, de modo a garantir equilíbrio de distribuição de dados. Um procedimento no qual um novo nó se une a um agrupamento se refere essencialmente à migração de dados, atualização de visualização após a migração de dados, notificação de atualização de visualização e processamento realizado por um nó de OSD relacionado na replicação de uma solicitação de IO de acordo com uma visualização atualizada.

[0255] O processamento de IO em um processo de migração de dados é consistente com o processamento em um procedimento de failback e um procedimento de recuperação de dados; após a migração de dados estar concluída, um processo no qual um MDC realiza atualização de visualização e notificação é consistente com o processamento de atualização de visualização realizado após um failback estar concluída. Um processo no qual cada nó de OSD relacionado processa, de acordo com a visualização atualizada, replicação de dados que correspondem à solicitação de IO é também consistente com o processamento de IO realizado por cada nó de OSD de acordo com uma última visualização após um failback estar concluída.

[0256] Em uma modalidade específica, realizar atualização de visualização após a migração de dados estar concluída pode incluir o seguinte: (1) O novo nó que se une ao agrupamento ainda é um nó de OSD secundário de alguns grupos de partição, uma situação de partição é consistente, e um OSD secundário original do grupo de partição não é mais um nó secundário da partição; (2) O novo nó que se une ao agrupamento é definido como um nó de OSD primário de alguns grupos de partição, e o grupo de partição não pertence mais a um nó de OSD primário original (o grupo de partição não é mais distribuído no OSD primário original).

[0257] A fim de facilitar o entendimento, a seção a seguir usa uma modalidade específica para descrição. Essa modalidade é executada pelo módulo de MDC, o módulo de roteamento de IO e o nó de OSD mencionado nas modalidades descritas na Figura 2A a Figura 2C e na Figura 5. A fim de facilitar o entendimento, é previsto que nessa modalidade, nós de OSD relacionados são um nó de OSD1, um nó de OSDn e um novo nó de OSD que se une a um agrupamento, em que o OSD1 é um nó de OSD primário de um grupo de partição P1 (P1 para abreviar) e um nó de OSD secundário de um grupo de partição Pn (Pn para abreviar), o OSDn é um nó de OSD primário do Pn e um nó de OSD secundário do P1, e o novo nó de OSD que se une ao agrupamento é um nó de OSD secundário do P1 e um nó de OSD secundário do Pn. Conforme mostrado na Figura 12A e na Figura 12B, o procedimento no qual o novo nó de OSD se une ao agrupamento nessa modalidade inclui especificamente as etapas a seguir:

[0258] 1202. Entregar um comando de configuração de que um novo nó de OSD se une a um agrupamento ao módulo de MDC.

[0259] Em um modo de implantação específico, um administrador de sistema pode notificar, usando-se o comando de configuração, o módulo de MDC de que o novo nó de OSD se une ao agrupamento.

[0260] 1202. O módulo de MDC realiza atualização de visualização, e migra partições em alguns nós de OSD para o novo nó de OSD que se une ao agrupamento.

[0261] Nessa modalidade, o módulo de MDC migra uma partição secundária do P1 no nó de OSD1 e uma partição secundária do Pn no nó de OSDn ao novo nó de OSD que se une ao agrupamento, de modo que o novo nó de OSD que se une ao agrupamento sirva como um novo nó de OSD secundário do P1 e um novo nó de OSD secundário do Pn.

[0262] 1204. Notificar o nó de OSD1 de atualização de visualização: adicionar um novo OSD secundário.

[0263] Em um modo de implantação específico, o módulo de MDC notifica o nó de OSD1 de atualização de visualização, o novo nó de OSD que se une ao agrupamento é adicionado a uma nova visualização de partição do P1 como o nó de OSD secundário do P1, e uma situação de partição correspondente é "inconsistente" (visto que o novo nó de OSD que se une ao

agrupamento e o nó de OSD1 não sincronizaram os dados do P1 ainda).

[0264] 1208. Notificar o nó de OSDn de atualização de visualização: adicionar um novo OSD secundário.

[0265] Em um modo de implantação específico, o novo nó de OSD que se une ao agrupamento é adicionado a uma nova visualização de partição do Pn como o nó de OSD secundário do Pn, e uma situação de partição correspondente é "inconsistente" (visto que o novo nó de OSD que se une ao agrupamento e o nó de OSDn não sincronizaram dados do Pn ainda).

[0266] 1210. Começar o novo nó de OSD que se une ao agrupamento.

[0267] Em um modo de implantação específico, após o novo nó de OSD que se une ao agrupamento se unir ao agrupamento como um novo nó de OSD, um processo de inicialização é realizado. Um processo específico é o mesmo que o procedimento anterior de geração inicializada e aquisição de uma visualização de agrupamento na Figura 6, e os detalhes não são descritos novamente no presente documento.

[0268] 1212. Retornar informações de partição do nó de OSD.

[0269] Em um modo de implantação específico, o módulo de MDC retorna uma visualização de uma partição no novo nó de OSD que se une ao agrupamento para o novo nó de OSD que se une ao agrupamento, ou seja, uma visualização de IO do P1 e uma visualização de IO do Pn nessa modalidade.

[0270] 1214. O novo nó de OSD que se une ao agrupamento solicita que o nó primário de OSD1 sincronize dados de partição.

[0271] Em um modo de implantação específico, o novo nó de OSD que se une ao agrupamento solicita, de acordo com a visualização de IO do P1 retornada pelo MDC, que o nó de OSD primário do P1, ou seja, que o OSD1, sincronize dados do P1, ou seja, sincronize dados de uma partição primária do P1 no nó de OSD1, de modo que os dados de uma partição secundária do P1 no novo nó de OSD que se une ao agrupamento sejam consistentes com os dados da partição primária do P1 no nó de OSD1.

[0272] Um procedimento de sincronização específica é similar a um procedimento de sincronização de dados de toda uma partição no procedimento de recuperação de dados anterior mostrado na Figura 10A e na Figura 10B, e os detalhes não são descritos novamente no presente documento.

[0273] 1216. O novo nó de OSD que se une ao agrupamento solicita que o nó primário de OSDn sincronize dados de partição.

[0274] Em um modo de implantação específico, o novo nó de OSD que se une ao agrupamento solicita, de acordo com a visualização de IO do Pn retornada pelo MDC, que o nó de OSD primário do Pn, ou seja, que o OSDn, sincronize dados do Pn, ou seja, sincronize dados de uma partição primária do Pn no nó de OSDn, de modo que os dados de uma partição secundária do Pn no novo nó de OSD que se une ao agrupamento sejam consistentes com os dados da partição primária do Pn no nó de OSDn.

[0275] Um procedimento de sincronização específica é similar a um procedimento de sincronização de dados de toda uma partição no procedimento de recuperação de dados anterior mostrado na Figura 10A e na Figura 10B, e os detalhes não são descritos novamente no presente documento.

[0276] 1218. Concluir sincronização de dados de partição.

[0277] 1220. Notificar o MDC que o nó secundário concluiu migração de dados de partição.

[0278] Em um modo de implantação específico, o nó de OSD1 notifica o módulo de MDC de que o novo nó de OSD que se une ao agrupamento concluiu a sincronização de dados do P1.

[0279] 1222. Notificar o MDC de que o nó secundário concluiu a migração de dados de partição.

[0280] Em um modo de implantação específico, o nó de OSDn notifica o módulo de MDC de que o novo nó de OSD que se une ao agrupamento concluiu a sincronização de dados do Pn.

[0281] 1224. O módulo de MDC realiza atualização de visualização.

[0282] Um princípio de atualização de visualização específico é o mesmo que aquele descrito nos procedimentos anteriores, um processo de atualização é o mesmo que aquele descrito nos procedimentos anteriores, e os detalhes não são descritos novamente no presente documento.

[0283] 1226 a 1230. Notificar separadamente o nó de OSD1, o nó de OSDn e o novo nó de OSD que se une ao agrupamento de atualização de visualização.

[0284] 1232. O nó de OSD1, o nó de OSDn e o novo nó de OSD que se une ao agrupamento processam, de acordo com uma visualização

atualizada, replicação de dados que correspondem a uma solicitação de IO.

[0285] Com base nas descrições anteriores das modalidades, uma pessoa de habilidade comum na técnica pode entender que todas ou algumas das etapas dos métodos nas modalidades podem ser implantadas por um programa que instrui hardware relevante. O programa pode ser armazenado em um meio de armazenamento legível por computador. Quando o programa é executado, as etapas das modalidades de método são realizadas. O meio de armazenamento é, por exemplo, uma ROM/RAM, um disco magnético ou um disco óptico.

[0286] As descrições anteriores são meramente modalidades específicas da presente invenção, porém, não se destinam a limitar o escopo de proteção da presente invenção. Qualquer variação ou substituição prontamente constatada por uma pessoa versada na técnica dentro do escopo técnico revelado na presente invenção deve estar dentro do escopo de proteção da presente invenção. Portanto, o escopo de proteção da presente invenção deve ser submetido ao escopo de proteção das reivindicações.

REIVINDICAÇÕES

1. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204) e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206), uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD (206) em que uma partição em um grupo de partição está localizada; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que

o módulo de MDC (202) determina que um nó de OSD (206) no sistema é um nó de OSD (206) defeituoso, determina uma partição no nó de OSD (206) defeituoso, atualiza uma visualização de partição de um grupo de partição que compreende a partição no nó de OSD (206) defeituoso, e envia uma notificação de atualização a um nó de OSD (206) primário na visualização de partição atualizada; e

o nó de OSD (206) primário é adaptado para processar, de acordo com a visualização de partição atualizada após o recebimento da notificação de

atualização enviada pelo módulo de MDC (202), replicação dos dados correspondentes à solicitação de IO

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD (206) primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD (206) primário; e

o processamento, de acordo com a visualização de partição atualizada, da replicação dos dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO de um módulo de roteamento de IO (204) em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD (206) secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que os dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

2. Sistema, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que o módulo de MDC (202) é adicionalmente adaptado para: gerar uma visualização de IO, em que a visualização de IO compreende um identificador de um nó de OSD (206) primário de um grupo de partição; e enviar a visualização de IO para o módulo de roteamento de IO (204) e para o nó de OSD (206) em que a partição na visualização de partição está localizada; e

o nó de OSD (206) primário é adicionalmente adaptado para: atualizar, de acordo com a visualização de partição atualizada, uma visualização de IO localmente armazenada no nó de OSD (206) primário, e processar, de acordo com a visualização de IO localmente armazenada atualizada, a replicação dos dados correspondentes à solicitação de IO.

3. Sistema, de acordo com a reivindicação 2, **CARACTERIZADO** pelo fato de que

o módulo de MDC (202) é adicionalmente adaptado para: ao determinar que a partição no nó de OSD (206) defeituoso compreende uma partição primária, atualizar uma visualização de IO de um grupo de partição que compreende a partição primária, e notificar um nó de OSD (206) secundário na visualização de partição atualizada da visualização de IO atualizada; e

o nó de OSD(206) secundário na visualização de partição atualizada é adaptado para: atualizar uma visualização de IO localmente armazenada de acordo com a visualização de IO atualizada, e processar, de acordo com a visualização de IO localmente armazenada atualizada, a replicação dos dados correspondentes à solicitação de IO.

4. Sistema, de acordo com a reivindicação 3, **CARACTERIZADO** pelo fato de que

o módulo de MDC (202) é adicionalmente adaptado para notificar o módulo de roteamento de IO (204) da visualização de IO atualizada; e

o módulo de roteamento de IO (204) atualiza, de acordo com a visualização de IO atualizada, uma visualização de IO localmente armazenada no módulo de roteamento de IO (204), e processa o encaminhamento da solicitação de IO de acordo com a visualização de IO localmente armazenada atualizada.

5. Sistema, de acordo com a reivindicação 3, **CARACTERIZADO** pelo fato de que a atualização de uma visualização de partição de um grupo de partição que compreende a partição no nó de OSD (206) defeituoso especificamente compreende:

quando a partição no nó de OSD (206) defeituoso compreender uma partição secundária, marcar uma situação de partição do nó de OSD (206) defeituoso em uma visualização de partição de um grupo de partição que compreenda a partição secundária como inconsistente; e

quando a partição no nó de OSD (206) defeituoso compreender a partição primária, definir o nó de OSD (206) defeituoso que serve como um nó de OSD (206) primário em uma visualização de partição do grupo de partição que compreenda a partição primária como um novo nó de OSD (206) secundário, marcar uma situação de partição correspondente ao novo nó de OSD (206)

secundário como inconsistente, selecionar um nó de OSD (206) secundário cuja situação de partição seja consistente com um nó de OSD (206) secundário original na visualização de partição do grupo de partição que compreende a partição primária, e definir o nó de OSD (206) secundário selecionado como um novo nó de OSD (206) primário.

6. Sistema, de acordo com a reivindicação 5, **CARACTERIZADO** pelo fato de que

o módulo de MDC (202) é adaptado para: após o nó de OSD (206) defeituoso completar failback e a recuperação de dados, atualizar adicionalmente a visualização de partição atualizada e a visualização de IO atualizada, enviar uma notificação de atualização a um nó de OSD (206) primário na visualização de partição atualizada adicional, e enviar uma notificação de atualização a um nó de OSD (206) secundário na visualização de partição atualizada adicional;

o nó de OSD (206) primário na visualização de partição atualizada adicional é adaptado para processar, de acordo com a visualização de partição atualizada adicional, a replicação dos dados correspondentes à solicitação de IO; e

o nó de OSD (206) secundário na visualização de partição atualizada adicional é adaptado para processar, de acordo com a visualização de IO atualizada adicional, a replicação dos dados correspondentes à solicitação de IO.

7. Sistema, de acordo com qualquer uma das reivindicações 1 a 6, **CARACTERIZADO** pelo fato de que o sistema compreende múltiplos hospedeiros, o módulo de MDC (202), os múltiplos módulos de roteamento de IO (204), e o nó de OSD (206) são separadamente instalados em pelo menos um dentre os múltiplos hospedeiros, e o nó de OSD (206) é adaptado para gerenciar recursos de armazenamento físico no hospedeiro.

8. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204), e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógica correspondente aos recursos de

armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206), uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um OSD primário de um grupo de partição; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar armazenamento de dados de IO de acordo com a solicitação de IO, **CARACTERIZADO** pelo fato de que

um módulo de roteamento de IO (204) é adaptado para: receber a solicitação de IO, em que a solicitação de IO compreende uma chave; determinar, de acordo com a chave, um grupo de partição ao qual os dados correspondentes à solicitação de IO pertencem, e determinar um nó de OSD (206) primário do grupo de partição ao qual os dados pertencem; adicionar informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem à solicitação de IO; e enviar, ao nó de OSD (206) primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO;

o nó de OSD (206) primário é adaptado para: armazenar a visualização de partição localmente no nó de OSD (206) primário, em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; receber a

solicitação de IO; executar a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD (206) primário; gerar uma solicitação de replicação que porta as informações de versão de visualização de IO; e enviar, de acordo com a visualização de partição localmente armazenada, a solicitação de replicação para um nó de OSD (206) secundário que é do grupo de partição ao qual os dados pertencem e cuja situação de partição é consistente na visualização de partição localmente armazenada, ou para um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada e um nó de OSD (206) secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada, mas que é de recuperação de dados; e

o nó de OSD (206) secundário é adaptado para: receber a solicitação de replicação, e executar a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD (206) secundário, de modo que os dados correspondentes à solicitação de IO no nó de OSD (206) secundário continuem consistentes com os dados correspondentes à solicitação de IO no nó de OSD (206) primário.

9. Sistema, de acordo com a reivindicação 8, **CARACTERIZADO** pelo fato de que o nó de OSD (206) primário é adicionalmente adaptado para: retornar um erro ao módulo de roteamento de IO (204) após determinar, de acordo com as informações de versão de visualização de IO, que a versão de visualização de IO na solicitação de IO é anterior à versão de visualização de IO localmente armazenada no OSD primário; e após determinar que a versão de visualização de IO na solicitação de IO é posterior à versão de visualização de IO localmente armazenada no OSD primário, adicionar a solicitação de IO a uma fila de cache, e consultar o módulo de MDC (202) quanto às informações de versão de visualização de IO da visualização de IO do grupo de partição ao qual os dados pertencem, de modo a executar a solicitação de IO após determinar que a versão de visualização de IO localmente armazenada no OSD primário é consistente com a versão de visualização de IO na solicitação de IO; e

o módulo de roteamento de IO (204) é adaptado para: após receber o erro retornado pelo nó de OSD (206) primário, consultar o módulo de MDC (202) quanto à visualização de IO do grupo de partição ao qual os dados pertencem e, após obter informações de versão de visualização de IO atualizada, enviar uma solicitação de IO que porta as informações de versão de visualização de IO atualizada.

10. Sistema, de acordo com a reivindicação 8, **CARACTERIZADO** pelo fato de que as informações de versão de visualização de IO compreendem um número de versão de visualização de IO, e o nó de OSD (206) primário gera adicionalmente um identificador de sequência para a solicitação de IO, e adiciona o identificador de sequência à solicitação de replicação enviada ao nó de OSD (206) secundário, em que o identificador de sequência compreende o número de versão de visualização de IO e um número de sequência, e o número de sequência indica um número de série de uma operação de modificação em dados correspondentes a um grupo de partição na visualização de IO dentro de uma versão de visualização de IO; e

o nó de OSD (206) secundário é adicionalmente adaptado para executar a solicitação de replicação de acordo com o identificador de sequência.

11. Sistema, de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que o nó de OSD (206) primário é adicionalmente adaptado para gravar uma entrada quando a solicitação de IO for uma operação de modificação, em que a entrada compreende um tipo de operação, um ID de grupo de partição do grupo de partição ao qual os dados pertencem, o identificador de sequência, e a chave.

12. Sistema, de acordo com a reivindicação 10, **CARACTERIZADO** pelo fato de que a solicitação de replicação porta adicionalmente um identificador de sequência em uma solicitação de replicação anterior enviada pelo nó de OSD (206) primário para o grupo de partição; e

o nó de OSD (206) secundário é adaptado para: após receber a solicitação de replicação, executar a solicitação de replicação quando o identificador de sequência na solicitação de replicação anterior for consistente com um maior identificador de sequência localmente armazenado no nó de OSD (206) secundário.

13. Sistema, de acordo com a reivindicação 12, **CARACTERIZADO**

pelo fato de que o nó de OSD (206) secundário é adicionalmente adaptado para: ao determinar que o identificador de sequência na solicitação de replicação anterior é maior do que o maior identificador de sequência localmente armazenado no nó de OSD (206) secundário, solicitar que o nó de OSD (206) primário reenvie uma solicitação faltante, de modo a manter os dados no nó de OSD (206) secundário consistentes com os dados no nó de OSD (206) primário.

14. Sistema, de acordo com a reivindicação 8, **CARACTERIZADO** pelo fato de que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que sejam de um OSD em que uma partição em um grupo de partição está localizada; e o módulo de MDC (202) é adicionalmente adaptado para:

ao detectar, em um processo de processamento da solicitação de IO, que o nó de OSD (206) primário se torna defeituoso, definir o nó de OSD (206) primário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD (206) secundário, e marcar uma situação de partição do novo OSD secundário como inconsistente; definir qualquer nó de OSD (206) secundário do nó de OSD (206) secundário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD (206) primário, notificar o novo nó de OSD (206) primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e atualizar, através do uso do novo nó de OSD (206) primário, a visualização de IO do grupo de partição ao qual os dados pertencem, e notificar o módulo de roteamento de IO (204) da visualização de IO atualizada da partição à qual os dados pertencem;

o módulo de roteamento de IO (204) é adicionalmente adaptado para: receber a visualização de IO atualizada que é do grupo de partição e que é enviada pelo módulo de MDC (202), e enviar a solicitação de IO para o novo nó de OSD (206) primário de acordo com a visualização de IO atualizada do grupo de partição; e

o novo nó de OSD (206) primário é adaptado para: receber a solicitação de IO e, após executar a solicitação de IO, gerar uma segunda solicitação de replicação, e enviar a segunda solicitação de replicação para um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição atualizada do grupo de partição ao qual os dados

pertencem.

15. Sistema, de acordo com a reivindicação 8, **CARACTERIZADO** pelo fato de que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um nó de OSD (206) em que uma partição em um grupo de partição está localizada; e o módulo de MDC (202) é adicionalmente adaptado para: ao detectar, em um processo de processamento da solicitação de IO, que qualquer nó de OSD (206) secundário do nó de OSD (206) secundário se torna defeituoso, marcar uma situação de partição de qualquer nó de OSD (206) secundário na visualização de partição do grupo de partição ao qual os dados pertencem como inconsistente, e notificar o nó de OSD (206) primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e

o nó de OSD (206) primário é adaptado para: após receber a visualização de partição atualizada do grupo de partição ao qual os dados pertencem, enviar a solicitação de replicação para um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição atualizada, e ignorar o envio da solicitação de replicação para o nó de OSD (206) secundário cuja situação de partição é inconsistente.

16. Sistema, de acordo com qualquer uma das reivindicações 8 a 15, **CARACTERIZADO** pelo fato de que o sistema compreende múltiplos hospedeiros, o módulo de MDC (202), os múltiplos módulos de roteamento de IO (204), e o nó de OSD (206) são separadamente instalados em pelo menos um dentre os múltiplos hospedeiros, e o nó de OSD (206) é adaptado para gerenciar recursos de armazenamento físico no hospedeiro.

17. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204), e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógicacorrespondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206), uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica

exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um OSD primário de um grupo de partição; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que

o nó de OSD (206) é adaptado para: enviar uma solicitação de consulta ao módulo de MDC (202) após um failback para solicitar uma visualização de IO de um grupo de partição que compreende uma partição no nó de OSD (206), em que o nó de OSD (206) é chamado de um nó de OSD (206) de failback, e a solicitação de consulta porta um identificador de OSD do nó de OSD (206) de failback; receber a visualização de IO retornada pelo módulo de MDC (202); iniciar uma solicitação de recuperação de dados para um OSD primário na visualização de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD (206) de failback durante uma falha; receber os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processar a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC (202) após o nó de OSD (206) de failback completar a recuperação de dados;

o módulo de MDC (202) é adaptado para: receber a solicitação de consulta do nó de OSD (206) de failback, retornar a visualização de IO ao nó de OSD (206) de failback de acordo com o identificador de OSD na solicitação de consulta, e atualizar a visualização de partição do grupo de partição após o nó

de OSD (206) de failback completar a recuperação de dados; e

o nó de OSD (206) primário é adaptado para: receber a solicitação de recuperação de dados do nó de OSD (206) de failback, enviar os dados atualizados durante a falha para o nó de OSD (206) de failback, e processar, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC (202) após o nó de OSD (206) de failback completar a recuperação de dados, a replicação de dados correspondentes à solicitação de IO;

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD (206) primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD (206) primário; e

o processamento, de acordo com a visualização de partição, da replicação de dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO de um módulo de roteamento de IO (204) em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD (206) secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

18. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que

o módulo de MDC (202) é adicionalmente adaptado para: após o nó de OSD (206) de failback se tornar defeituoso e antes do failback, atualizar a visualização de partição e a visualização de IO que são do grupo de partição; e

o retorno da visualização de IO para o nó de OSD (206) de failback de acordo com o identificador de OSD na solicitação de consulta é especificamente retornar a visualização de IO atualizada para o nó de OSD (206) de failback.

19. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que

o nó de OSD (206) primário é adicionalmente adaptado para: após receber a solicitação de recuperação de dados, receber a solicitação de IO enviada por um módulo de roteamento de IO (204) para a partição no nó de OSD (206) de failback, executar a solicitação de IO, e enviar, para o nó de OSD (206) de failback, uma solicitação de replicação que porta informações de chave de IO e os dados correspondentes à solicitação de IO; e

o nó de OSD (206) de failback grava, em um registro, a solicitação de replicação que porta as informações de chave de IO e os dados correspondentes à solicitação de IO, e atualiza, de acordo com uma gravação do registro após a recuperação de dados ser concluída, os dados correspondentes à solicitação de IO para recursos de armazenamento físico gerenciados pelo nó de OSD (206) de failback.

20. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que a solicitação de recuperação de dados porta um maior identificador de sequência que é de uma operação de IO para a partição no nó de OSD (206) de failback e que é localmente gravada no nó de OSD (206) de failback, em que o maior identificador de sequência é um último número de versão de visualização de IO da visualização de IO do grupo de partição que compreende a partição no nó de OSD (206) de failback, e um maior número de série de uma operação de modificação em dados correspondentes a uma partição na visualização de IO correspondente ao último número de versão de visualização de IO; e

o envio dos dados atualizados durante a falha para o nó de OSD (206) de failback compreende:

determinar que o maior identificador de sequência na solicitação de recuperação de dados é maior ou igual a um menor identificador de sequência atual localmente armazenado no nó de OSD (206) primário, enviar uma entrada que o nó de OSD (206) de failback não tem durante a falha para o nó de OSD

(206) de failback, receber uma solicitação de recuperação de dados iniciada pelo nó de OSD (206) de failback de acordo com a entrada, e enviar dados correspondentes à entrada para o nó de OSD (206) de failback, em que

o menor identificador de sequência é um menor número de versão de visualização de IO que está na visualização de IO do grupo de partição e que é armazenado no nó de OSD (206) primário, e um menor número de série de uma operação de modificação em dados correspondentes a uma partição na visualização de IO corresponde ao menor número de versão de visualização de IO.

21. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que a solicitação de recuperação de dados porta um maior identificador de sequência que é de uma operação de IO para a partição no nó de OSD(206) de failback e que é localmente gravado no nó de OSD (206) de failback, em que

o maior identificador de sequência compreende: um último número de versão de visualização de IO da visualização de IO do grupo de partição que compreende a partição no nó de OSD (206) de failback, e um maior número de série de uma operação de modificação em dados correspondentes a uma partição na visualização de IO dentro da visualização de IO corresponde ao último número de versão de visualização de IO; e

o envio dos dados atualizados durante a falha para o nó de OSD de (206) failback compreende:

determinar que o maior identificador de sequência na solicitação de recuperação de dados é menor do que um menor identificador de sequência atual localmente armazenado no nó de OSD (206) primário, enviar o menor identificador de sequência atual localmente armazenado no OSD primário para o nó de OSD (206) de failback, receber uma solicitação de recuperação de dados, iniciada pelo nó de OSD (206) de failback, para sincronizar todos os dados correspondentes a uma partição primária que pertence ao grupo de partição e que está no nó de OSD (206) primário, e enviar todos os dados correspondentes à partição primária para o nó de OSD (206) de failback, em que

o menor identificador de sequência é um menor número de versão de visualização de IO que está na visualização de IO do grupo de partição e que é armazenado no nó de OSD (206) primário, e um menor número de série de uma

operação de modificação em dados correspondentes a uma partição na visualização de IO corresponde ao menor número de versão de visualização de IO.

22. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que o módulo de MDC (202) é adicionalmente adaptado para: antes de atualizar a visualização de partição do grupo de partição após o nó de OSD (206) de failback completar a recuperação de dados, adaptado para

receber uma notificação de solicitação de atualização de visualização enviada pelo nó de OSD (206) primário, em que a notificação de solicitação de atualização de visualização porta um último número de versão de visualização de partição que é do grupo de partição e que está localmente armazenado no nó de OSD (206) primário; e executar a operação de atualização após determinar que o último número de versão de visualização de partição que é do grupo de partição e que está localmente armazenado no nó de OSD (206) primário é consistente com um último número de versão de visualização de partição localmente mantido pelo módulo de MDC (202).

23. Sistema, de acordo com a reivindicação 17, **CARACTERIZADO** pelo fato de que a atualização da visualização de partição do grupo de partição especificamente compreende:

atualizar uma situação de partição do nó de OSD (206) de failback na visualização de partição para consistente e, após determinar que o nó de OSD (206) de failback é um nó de OSD (206) primário do grupo de partição antes da falha, redefinir o nó de OSD (206) de failback como um novo nó de OSD (206) primário do grupo de partição, e definir o nó de OSD (206) primário como um novo nó de OSD (206) secundário do grupo de partição.

24. Sistema, de acordo com qualquer uma das reivindicações 17 a 23, **CARACTERIZADO** pelo fato de que o sistema compreende múltiplos hospedeiros, o módulo de MDC (202), os múltiplos módulos de roteamento de IO (204), e o nó de OSD (206) são separadamente instalados em pelo menos um dentre os múltiplos hospedeiros, e o nó de OSD (206) é adaptado para gerenciar recursos de armazenamento físico no hospedeiro.

25. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204), e múltiplos nós de

dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206), uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição são localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que

o sistema compreende uma memória (502) e um processador (506), em que

a memória (502) é adaptada para armazenar uma instrução legível por computador, e a instrução é usada para executar funções do módulo de MDC (202), do módulo de roteamento de IO (204), e do nó de OSD (206); e

o processador (506) é adaptado para: ser conectado à memória (502), ler a instrução na memória (502), e possibilitar, de acordo com a instrução, que o processador (506) execute as seguintes operações:

determinar que um nó de OSD (206) no sistema é um nó de OSD (206) defeituoso, determinar uma partição no nó de OSD (206) defeituoso, atualizar uma visualização de partição de um grupo de partição que compreende a partição no nó de OSD (206) defeituoso, e enviar uma notificação de

atualização para um nó de OSD (206) primário em que o grupo de partição na visualização de partição atualizada está localizado, de modo que o nó de OSD (206) primário processe, de acordo com a visualização de partição atualizada, a replicação de dados correspondentes à solicitação de IO;

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD (206) primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD (206) primário; e

o processamento, de acordo com a visualização de partição atualizada, de replicação dos dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO de um módulo de roteamento de IO (204) em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD (206) secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

26. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204), e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206), uma réplica é selecionada a partir das múltiplas réplicas da partição como uma

réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um OSD primário de um grupo de partição; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que

o sistema compreende uma memória (502) e um processador (506), em que

a memória (502) é adaptada para armazenar uma instrução legível por computador, e a instrução é usada para executar funções do módulo de MDC (202), do módulo de roteamento de IO (204), e do nó de OSD (206); e

o processador (506) é adaptado para: ser conectado à memória (502), ler a instrução na memória (502), e possibilitar, de acordo com a instrução, que o processador (506) execute as seguintes operações:

permitir que o módulo de roteamento de IO (204): receba a solicitação de IO, em que a solicitação de IO compreende uma chave; determine, de acordo com a chave, um grupo de partição ao qual os dados correspondentes à solicitação de IO pertencem, e determine um nó de OSD (206) primário do grupo de partição ao qual os dados pertencem; adicione informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem à solicitação de IO; e envie, ao nó de OSD (206) primário determinado, a solicitação de IO que porta as informações de versão de

visualização de IO;

permitir que o nó de OSD (206) primário: armazene a visualização de partição localmente no nó de OSD (206) primário, em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; receba a solicitação de IO; execute a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO localmente armazenada; gere uma solicitação de replicação que porta as informações de versão de visualização de IO; e envie, de acordo com a visualização de partição localmente armazenada, a solicitação de replicação para um nó de OSD (206) secundário que é da partição à qual os dados pertencem e cuja situação de partição é consistente na visualização de partição localmente armazenada, ou para um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada e um nó de OSD (206) secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada, mas que é de recuperação de dados; e

permitir que o nó de OSD (206) secundário: receba a solicitação de replicação, e execute a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD (206) secundário, de modo que os dados correspondentes à solicitação de IO no nó de OSD (206) secundário continuem consistentes com os dados correspondentes à solicitação de IO no nó de OSD (206) primário.

27. Sistema de replicação e armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC) (202), múltiplos módulos de roteamento de IO (204), e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD) (206), em que o módulo de MDC (202) é adaptado para: configurar, para cada nó de OSD (206), pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD (206), em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD (206),

uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD (206), um nó de OSD (206) em que a partição primária está localizada é um nó de OSD (206) primário de um grupo de partição que compreende a partição primária, e um nó de OSD (206) em que a partição secundária está localizada é um nó de OSD (206) secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um OSD primário de um grupo de partição; cada módulo de roteamento de IO (204) é adaptado para rotear uma solicitação de IO recebida para um nó de OSD (206); e o nó de OSD (206) é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que

o sistema compreende uma memória (502) e um processador (506), em que

a memória (502) é adaptada para armazenar uma instrução legível por computador, e a instrução é usada para executar funções do módulo de MDC (202), do módulo de roteamento de IO (204), e do nó de OSD (206); e

o processador (506) é adaptado para: ser conectado à memória (502), ler a instrução na memória (502), e possibilitar, de acordo com a instrução, que o processador (506) execute as seguintes operações:

permitir que o nó de OSD (206): envie uma solicitação de consulta ao módulo de MDC (202) após um failback para solicitar uma visualização de IO de um grupo de partição que compreende uma partição no nó de OSD (206), em que o nó de OSD (206) é chamado de um nó de OSD (206)de failback, e a solicitação de consulta porta um identificador de OSD do nó de OSD (206) de failback; receba a visualização de IO retornada pelo módulo de MDC (202); inicie uma solicitação de recuperação de dados para um OSD primário na visualização

de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD (206) de failback durante uma falha; receba os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processe a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC (202) após o nó de OSD (206) de failback completar a recuperação de dados;

permitir que o módulo de MDC (202): receba a solicitação de consulta do nó de OSD (206) de failback, retorne a visualização de IO ao nó de OSD (206) de failback de acordo com o identificador de OSD na solicitação de consulta, e atualize a visualização de partição do grupo de partição após o nó de OSD (206) de failback completar a recuperação de dados; e

permitir que o nó de OSD (206) primário: receba a solicitação de recuperação de dados do nó de OSD (206) de failback, envie os dados atualizados durante a falha para o nó de OSD (206)de failback, e processe, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC (202) após o nó de OSD (206) de failback completar a recuperação de dados, a replicação de dados correspondentes à solicitação de IO;

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD (206) primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD (206) primário; e

o processamento, de acordo com a visualização de partição, de replicação dos dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO de um módulo de roteamento de IO (204) em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD (206) secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD (206) secundário cuja situação de partição é

inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

28. Método para gerenciar armazenamento e replicação de dados em um sistema de armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO, e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para: configurar, para cada nó de OSD, pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD, em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD, uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD, um nó de OSD em que a partição primária está localizada é um nó de OSD primário de um grupo de partição que compreende a partição primária, e um nó de OSD em que a partição secundária está localizada é um nó de OSD secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada; cada módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que o método compreende:

determinar que um nó de OSD no sistema é um nó de OSD defeituoso, determinar uma partição no nó de OSD defeituoso, atualizar uma visualização de partição de um grupo de partição que compreende a partição no nó de OSD defeituoso, e enviar uma notificação de atualização para um nó de

OSD primário na visualização de partição atualizada; e

processar, pelo nó de OSD primário, de acordo com a visualização de partição atualizada após o recebimento da notificação de atualização enviada pelo módulo de MDC, replicação dos dados correspondentes à solicitação de IO;

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD primário; e

o processamento, de acordo com a visualização de partição atualizada, da replicação dos dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO do módulo de roteamento de IO em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um nó de OSD secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

29. Método, de acordo com a reivindicação 28, **CARACTERIZADO** pelo fato de que o método compreende adicionalmente:

ao determinar que a partição no nó de OSD defeituoso compreende uma partição primária, atualizar, através do módulo de MDC, uma visualização de IO de um grupo de partição que compreende a partição primária, e notificar um nó de OSD secundário na visualização de partição atualizada da visualização de IO atualizada; e

atualizar, através do nó de OSD secundário na visualização de partição atualizada, uma visualização de IO localmente armazenada de acordo com a visualização de IO atualizada, e processar, de acordo com a visualização

de IO localmente armazenada atualizada, a replicação dos dados correspondentes à solicitação de IO.

30. Método, de acordo com a reivindicação 29, **CARACTERIZADO** pelo fato de que a atualização de uma visualização de partição de um grupo de partição que compreende a partição no nó de OSD (206) defeituoso especificamente compreende:

quando a partição no nó de OSD defeituoso compreender uma partição secundária, marcar uma situação de partição do nó de OSD defeituoso em uma visualização de partição de um grupo de partição que compreenda a partição secundária como inconsistente; e

quando a partição no nó de OSD defeituoso compreender a partição primária, definir o nó de OSD defeituoso que serve como um nó de OSD primário em uma visualização de partição do grupo de partição que compreenda a partição primária como um novo nó de OSD secundário, marcar uma situação de partição correspondente ao novo nó de OSD secundário como inconsistente, selecionar um nó de OSD secundário cuja situação de partição seja consistente com um nó de OSD secundário original na visualização de partição do grupo de partição que compreende a partição primária, e definir o nó de OSD secundário selecionado como um novo nó de OSD primário.

31. Método, de acordo com a reivindicação 30, **CARACTERIZADO** pelo fato de que o método compreende adicionalmente:

após o nó de OSD defeituoso completar o failback e recuperação de dados, atualizar adicionalmente, através do módulo de MDC, a visualização de partição atualizada e a visualização de IO atualizada, enviar uma notificação de atualização para um nó de OSD primário na visualização de partição atualizada adicional, e enviar uma notificação de atualização para um nó de OSD secundário na visualização de partição atualizada adicional;

processar, através do nó de OSD primário na visualização de partição atualizada adicional de acordo com a visualização de partição atualizada adicional, a replicação dos dados correspondentes à solicitação de IO; e

processar, através do nó de OSD secundário na visualização de partição atualizada adicional de acordo com a visualização de IO atualizada adicional, a replicação dos dados correspondentes à solicitação de IO.

32. Método para gerenciar armazenamento e replicação de dados em

um sistema de armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO, e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para: configurar, para cada nó de OSD, pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD, em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD , uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD, um nó de OSD em que a partição primária está localizada é um nó de OSD primário de um grupo de partição que compreende a partição primária, e um nó de OSD em que a partição secundária está localizada é um nó de OSD secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um nó de OSD primário de um grupo de partição; cada módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar armazenamento de dados de IO de acordo com a solicitação de IO, **CARACTERIZADO** pelo fato de que o método compreende:

um módulo de roteamento de IO ser adaptado para: receber a solicitação de IO, em que a solicitação de IO compreende uma chave; determinar, de acordo com a chave, um grupo de partição ao qual os dados correspondentes à solicitação de IO pertencem, e determinar um nó de OSD primário do grupo de partição ao qual os dados pertencem; adicionar informações de versão de visualização de IO de uma visualização de IO do grupo de partição ao qual os dados pertencem à solicitação de IO; e enviar, para nó de OSD primário determinado, a solicitação de IO que porta as informações de versão de visualização de IO;

o nó de OSD primário ser adaptado para: armazenar a visualização de partição localmente no nó de OSD primário, em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; receber a solicitação de IO; executar a solicitação de IO após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de IO é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD primário; gerar uma solicitação de replicação que porta as informações de versão de visualização de IO; e enviar, de acordo com a visualização de partição localmente armazenada, a solicitação de replicação para um nó de OSD secundário que é do grupo de partição ao qual os dados pertencem e cuja situação de partição é consistente na visualização de partição localmente armazenada, ou para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada e um nó de OSD secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada, mas que é de recuperação de dados; e

o nó de OSD secundário ser adaptado para: receber a solicitação de replicação, e executar a solicitação de replicação após determinar, de acordo com as informações de versão de visualização de IO, que uma versão de visualização de IO na solicitação de replicação é consistente com uma versão de visualização de IO localmente armazenada no nó de OSD secundário, de modo que dados correspondentes à solicitação de IO no nó de OSD secundário continuem consistentes com dados correspondentes à solicitação de IO no nó de OSD primário.

33. Método, de acordo com a reivindicação 32, **CARACTERIZADO** pelo fato de que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que sejam de um OSD em que uma partição em um grupo de partição está localizada; e o método compreende adicionalmente:

ao detectar, em um processo de processamento da solicitação de IO, que o nó de OSD primário se torna defeituoso, definir, através do módulo de MDC, o nó de OSD primário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD secundário, e marcar uma

situação de partição do novo OSD secundário como inconsistente; definir qualquer nó de OSD secundário do nó de OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como um novo nó de OSD primário, notificar o novo nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e atualizar, através do uso do novo nó de OSD primário, a visualização de IO do grupo de partição ao qual os dados pertencem, e notificar o módulo de roteamento de IO da visualização de IO atualizada da partição à qual os dados pertencem;

o módulo de roteamento de IO é adicionalmente adaptado para: receber a visualização de IO atualizada que é do grupo de partição e que é enviada pelo módulo de MDC, e enviar a solicitação de IO para o novo nó de OSD primário de acordo com a visualização de IO atualizada do grupo de partição; e

o novo nó de OSD primário é adaptado para: receber a solicitação de IO e, após executar a solicitação de IO, gerar uma segunda solicitação de replicação, e enviar a segunda solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada do grupo de partição ao qual os dados pertencem.

34. Método, de acordo com a reivindicação 32, **CARACTERIZADO** pelo fato de que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que sejam de um OSD em que uma partição em um grupo de partição está localizada; e o método compreende adicionalmente:

ao detectar, em um processo de processamento da solicitação de IO, que qualquer nó de OSD secundário do nó de OSD secundário se torna defeituoso, marcar, através do módulo de MDC, uma situação de partição de qualquer nó de OSD secundário na visualização de partição do grupo de partição ao qual os dados pertencem como inconsistente, e notificar o nó de OSD primário da visualização de partição atualizada do grupo de partição ao qual os dados pertencem; e

o nó de OSD primário é adaptado para: após receber a visualização de partição atualizada do grupo de partição ao qual os dados pertencem, enviar a solicitação de replicação para um nó de OSD secundário cuja situação de partição é consistente na visualização de partição atualizada, e ignorar o envio

da solicitação de replicação para o nó de OSD secundário cuja situação de partição é inconsistente.

35. Método para gerenciar armazenamento e replicação de dados em um sistema de armazenamento distribuído, em que o sistema compreende pelo menos um módulo de controle de metadados (MDC), múltiplos módulos de roteamento de IO, e múltiplos nós de dispositivo de armazenamento com base em objeto (OSD), em que o módulo de MDC é adaptado para: configurar, para cada nó de OSD, pelo menos uma partição (partition) lógica correspondente aos recursos de armazenamento físico gerenciados por cada nó de OSD, em que uma partição tem múltiplas réplicas armazenadas em diferentes nós de OSD , uma réplica é selecionada a partir das múltiplas réplicas da partição como uma réplica primária, e a réplica primária é chamada de partição primária, uma réplica exceto a réplica primária da partição é chamada de partição secundária, a partição primária e a partição secundária correspondente à partição primária constituem um grupo de partição, a partição primária e a partição secundária em um mesmo grupo de partição estão localizadas em diferentes nós de OSD, um nó de OSD em que a partição primária está localizada é um nó de OSD primário de um grupo de partição que compreende a partição primária, e um nó de OSD em que a partição secundária está localizada é um nó de OSD secundário de um grupo de partição que compreende a partição secundária; e gerar uma visualização de partição e uma visualização de IO de acordo com a partição, em que a visualização de partição compreende informações a respeito de um OSD em que uma partição em um grupo de partição está localizada, e a visualização de IO compreende um identificador de um OSD primário de um grupo de partição; cada módulo de roteamento de IO é adaptado para rotear uma solicitação de IO recebida para um nó de OSD; e o nó de OSD é adaptado para executar, de acordo com a solicitação de IO, o armazenamento de dados correspondentes à solicitação de IO, **CARACTERIZADO** pelo fato de que o método compreende:

o nó de OSD é adaptado para: enviar uma solicitação de consulta ao módulo de MDC após um failback para solicitar uma visualização de IO de um grupo de partição que compreende uma partição no nó de OSD, em que o nó de OSD é chamado de um nó de OSD de failback, e a solicitação de consulta porta um identificador de OSD do nó de OSD de failback; receber a visualização de IO

retornada pelo módulo de MDC; iniciar uma solicitação de recuperação de dados para um OSD primário na visualização de IO, para solicitar a recuperação de dados atualizados pelo nó de OSD de failback durante uma falha; receber os dados que são atualizados durante a falha e que são enviados pelo OSD primário; e processar a replicação da solicitação de IO de acordo com uma visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback completar a recuperação de dados;

o módulo de MDC é adaptado para: receber a solicitação de consulta do nó de OSD de failback, retornar a visualização de IO para o nó de OSD de failback de acordo com o identificador de OSD na solicitação de consulta, e atualizar a visualização de partição do grupo de partição após o nó de OSD de failback completar a recuperação de dados; e

o nó de OSD primário é adaptado para: receber a solicitação de recuperação de dados do nó de OSD de failback, enviar os dados atualizados durante a falha para o nó de OSD de failback, e processar, de acordo com a visualização de partição que é do grupo de partição e que é atualizada pelo módulo de MDC após o nó de OSD de failback completar a recuperação de dados, a replicação de dados correspondentes à solicitação de IO;

em que a visualização de partição especificamente compreende uma identidade primária/secundária e uma situação de partição correspondente que são de um OSD em que uma partição em um grupo de partição está localizada; e o nó de OSD primário é adicionalmente adaptado para atualizar, de acordo com a visualização de partição atualizada, uma visualização de partição localmente armazenada no nó de OSD primário; e

o processamento, de acordo com a visualização de partição, de replicação de dados correspondentes à solicitação de IO especificamente compreende:

replicar, de acordo com a visualização de partição localmente armazenada atualizada, os dados correspondentes à solicitação de IO de um módulo de roteamento de IO em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada, ou em um nó de OSD secundário cuja situação de partição é consistente na visualização de partição localmente armazenada atualizada e um

nó de OSD secundário cuja situação de partição é inconsistente na visualização de partição localmente armazenada atualizada, mas que é de recuperação de dados, em que “consistente” indica que dados na partição secundária são consistentes com dados na partição primária, e “inconsistente” indica que dados na partição secundária são inconsistentes com dados na partição primária.

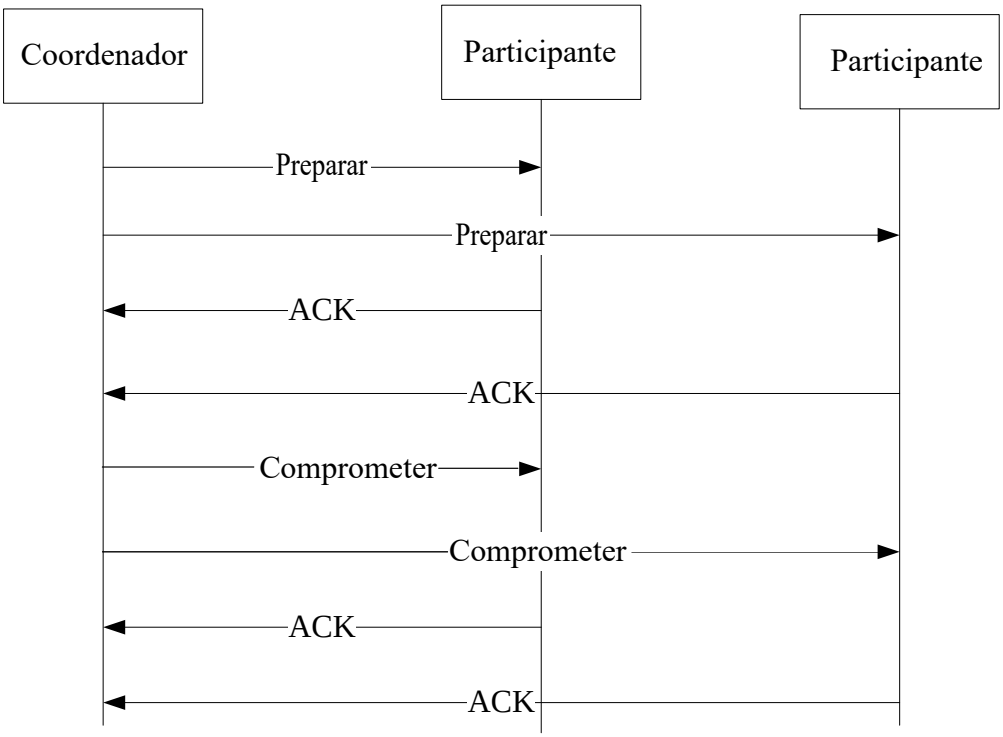

Coordenador
Participante
Participante
Preparar
Preparar
ACK
ACK
Comprometer
Comprometer
ACK
ACK

FIG. 1

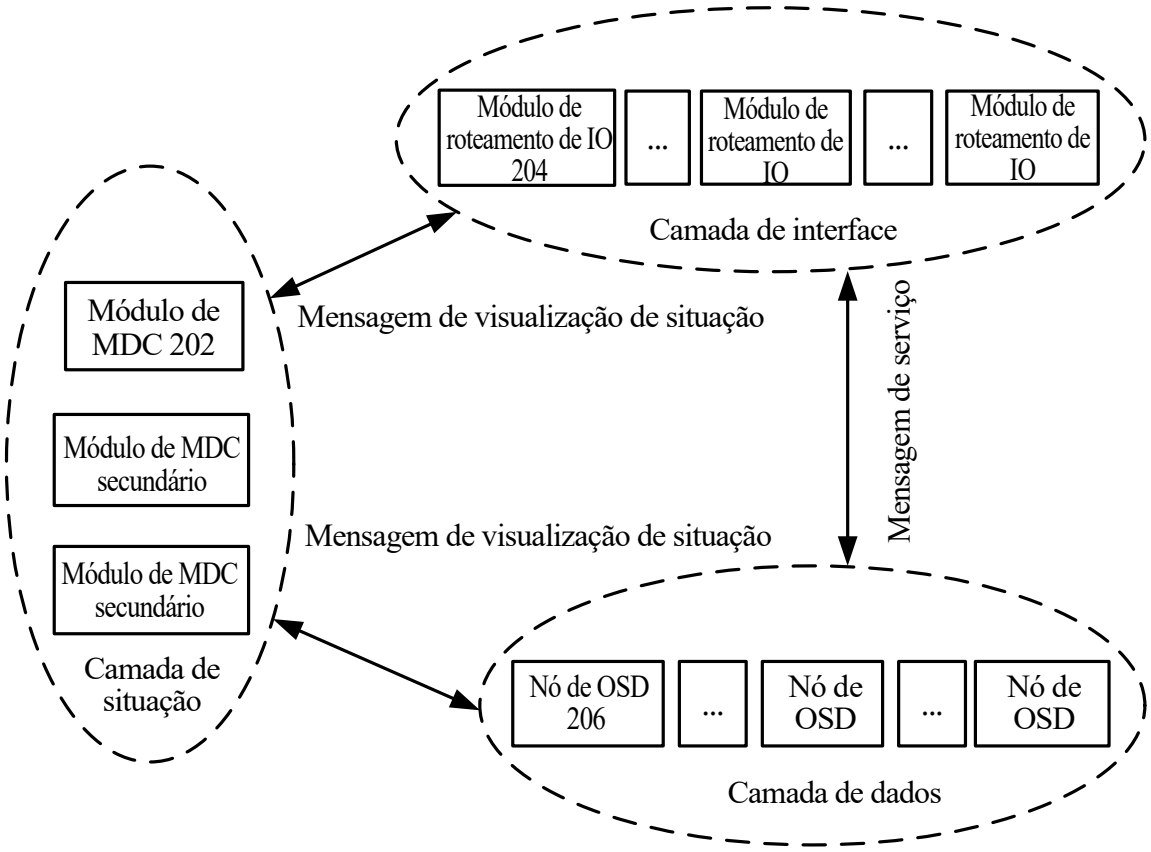

Módulo de roteamento de IO 204
...
Módulo de roteamento de IO
...
Módulo de roteamento de IO
Camada de interface
Módulo de MDC 202
Módulo de MDC secundário
Módulo de MDC secundário
Camada de situação
Mensagem de visualização de situação
Mensagem de visualização de situação
Mensagem de serviço
Nó de OSD 206
...
Nó de OSD
...
Nó de OSD
Camada de dados

FIG. 2A

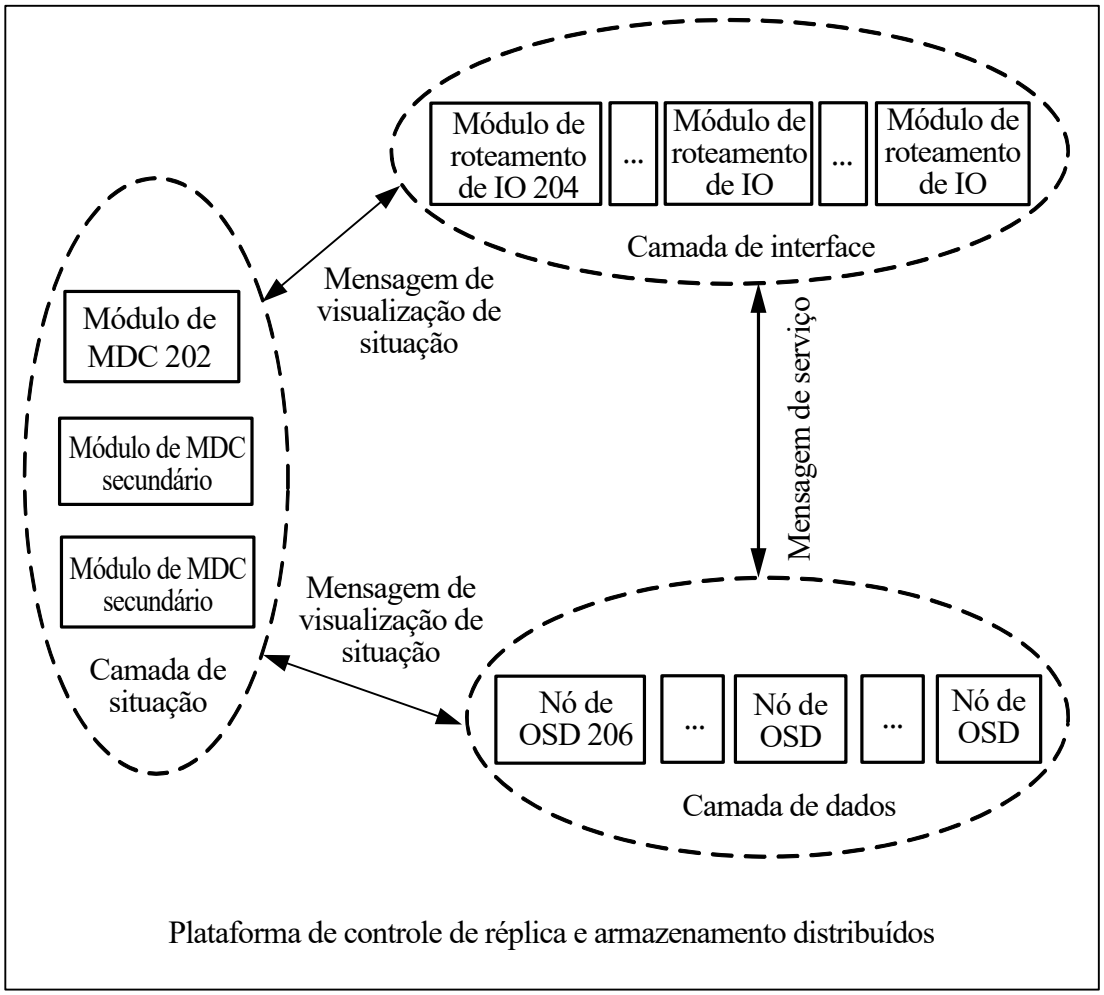

Módulo de roteamento de IO 204
...
Módulo de roteamento de IO
...
Módulo de roteamento de IO
Camada de interface
Mensagem de visualização de situação
Mensagem de serviço
Módulo de MDC 202
Módulo de MDC secundário
Módulo de MDC secundário
Camada de situação
Mensagem de visualização de situação
Nó de OSD 206
...
Nó de OSD
...
Nó de OSD
Camada de dados
Plataforma de controle de réplica e armazenamento distribuídos

FIG. 2B

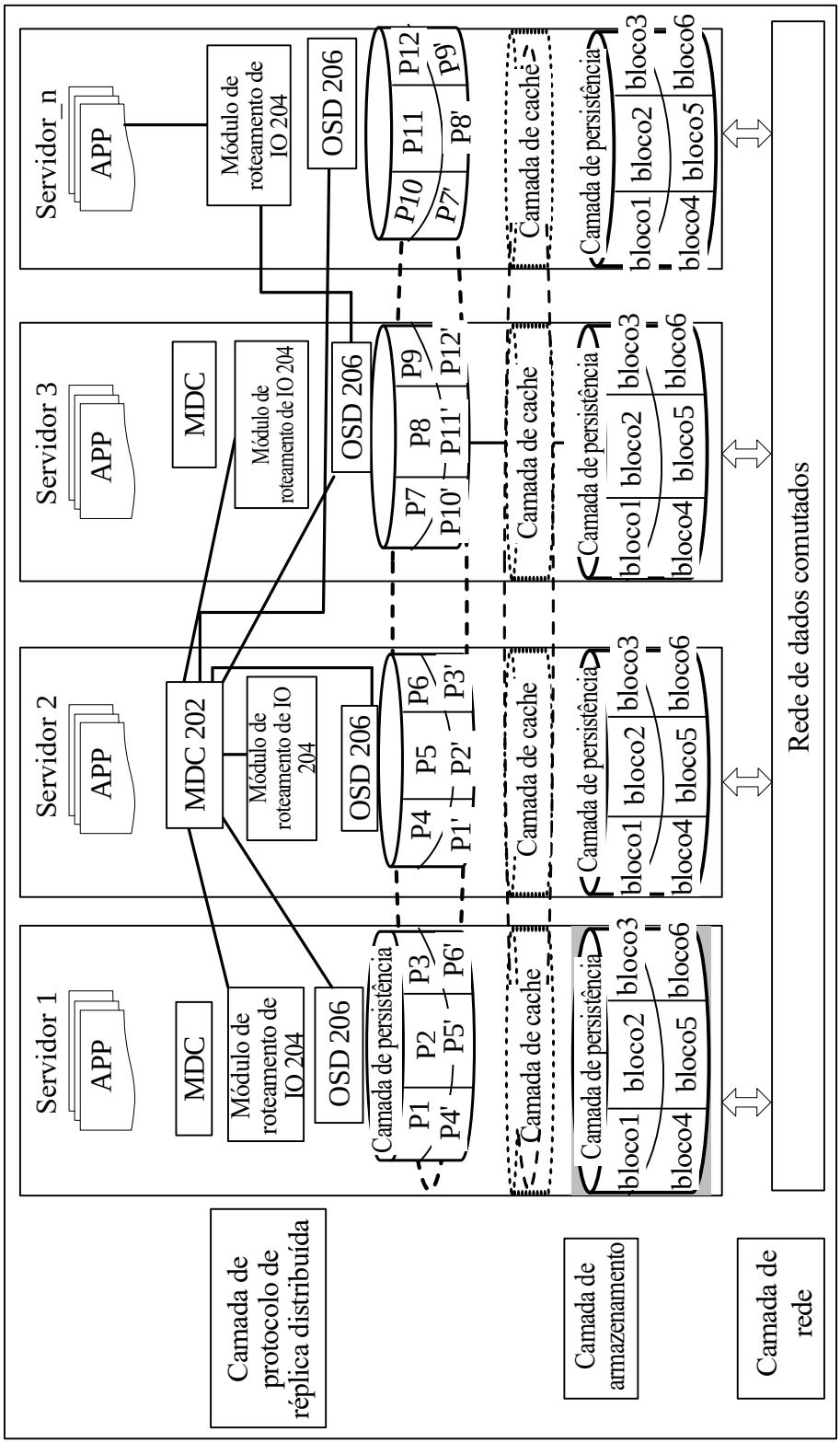

Servidor 1
APP
MDC
Módulo de roteamento de IO 204
OSD 206
Camada de persistência
P1
P4'
P2
P5'
P3
P6'
Camada de cache
bloco1
bloco2
bloco3
bloco4
bloco5
bloco6
Servidor 2
APP
MDC 202
Módulo de roteamento de IO 204
OSD 206
P4
P1'
P5
P2'
P6
P3'
Camada de cache
Camada de persistência
Servidor 3
APP
MDC
Módulo de roteamento de IO 204
OSD 206
P7
P10'
P8
P11'
P9
P12'
Camada de cache
Camada de persistência
Servidor_n
APP
Módulo de roteamento de IO 204
OSD 206
P10
P7'
P11
P8'
P12
P9'
Camada de cache
Camada de persistência
Rede de dados comutados
Camada de protocolo de réplica distribuída
Camada de armazenamento
Camada de rede

FIG. 2C

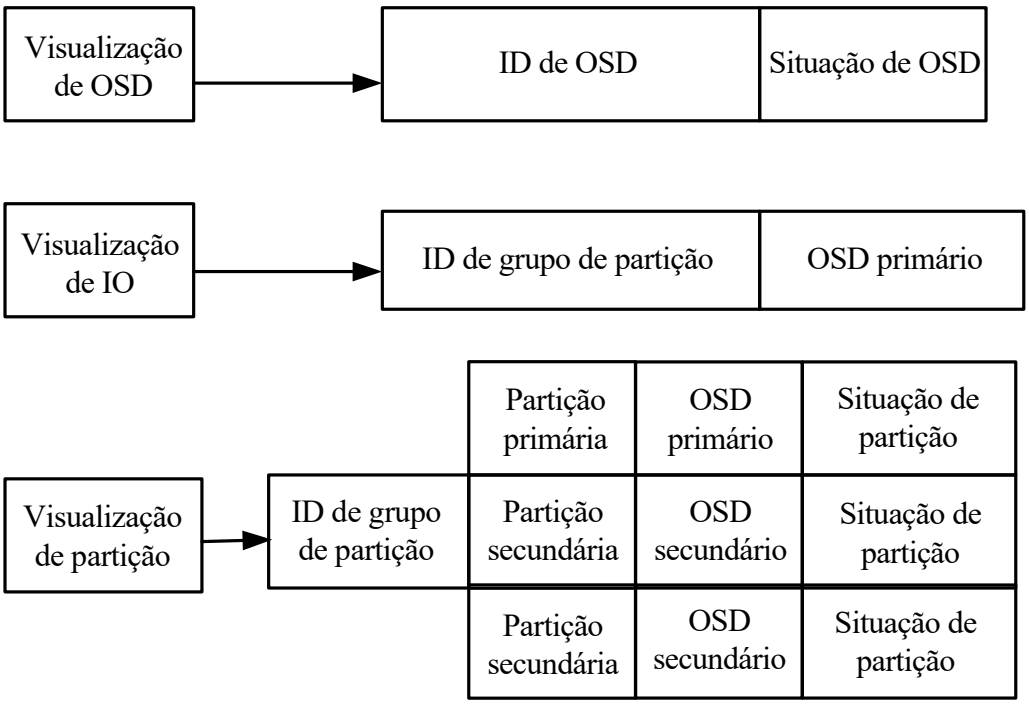

Visualização de OSD
ID de OSD
Situação de OSD
Visualização de IO
ID de grupo de partição
OSD primário
Visualização de partição
ID de grupo de partição
Partição primária
OSD primário
Situação de partição
Partição secundária
OSD secundário
Situação de partição
Partição secundária
OSD secundário
Situação de partição

FIG. 3

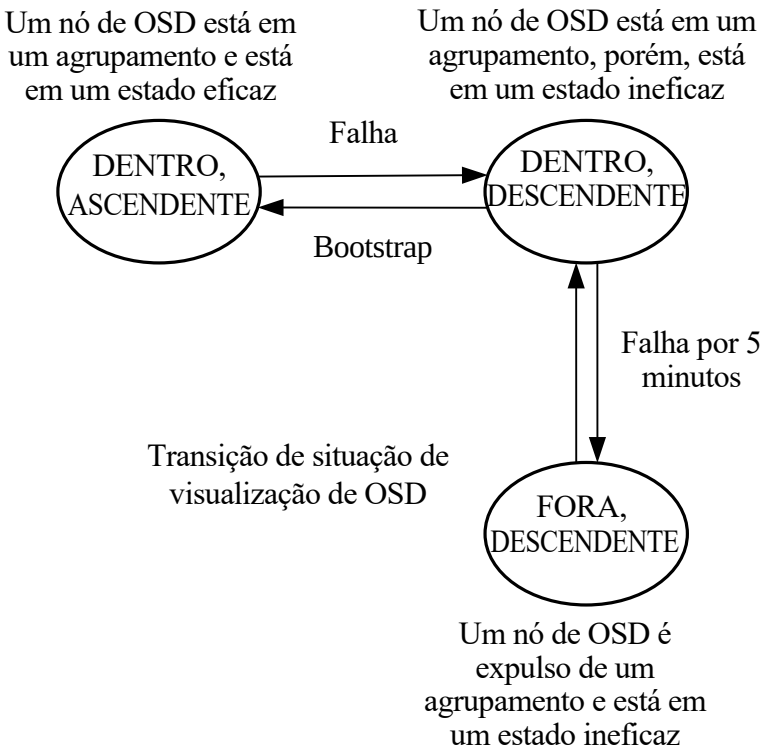

Um nó de OSD está em um agrupamento e está em um estado eficaz
Um nó de OSD está em um agrupamento, porém, está em um estado ineficaz
Falha
DENTRO, ASCENDENTE
DENTRO, DESCENDENTE
Bootstrap
Falha por 5 minutos
Transição de situação de visualização de OSD
FORA, DESCENDENTE
Um nó de OSD é expulso de um agrupamento e está em um estado ineficaz

FIG. 4

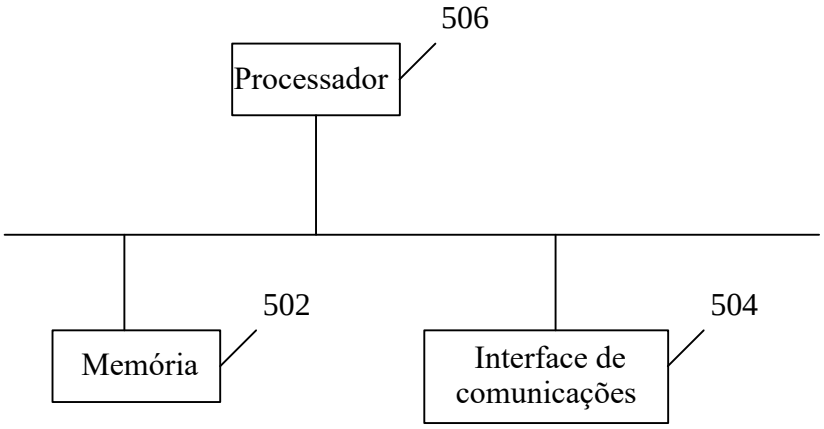

506
Processador
502
Memória
504
Interface de comunicações

FIG. 5

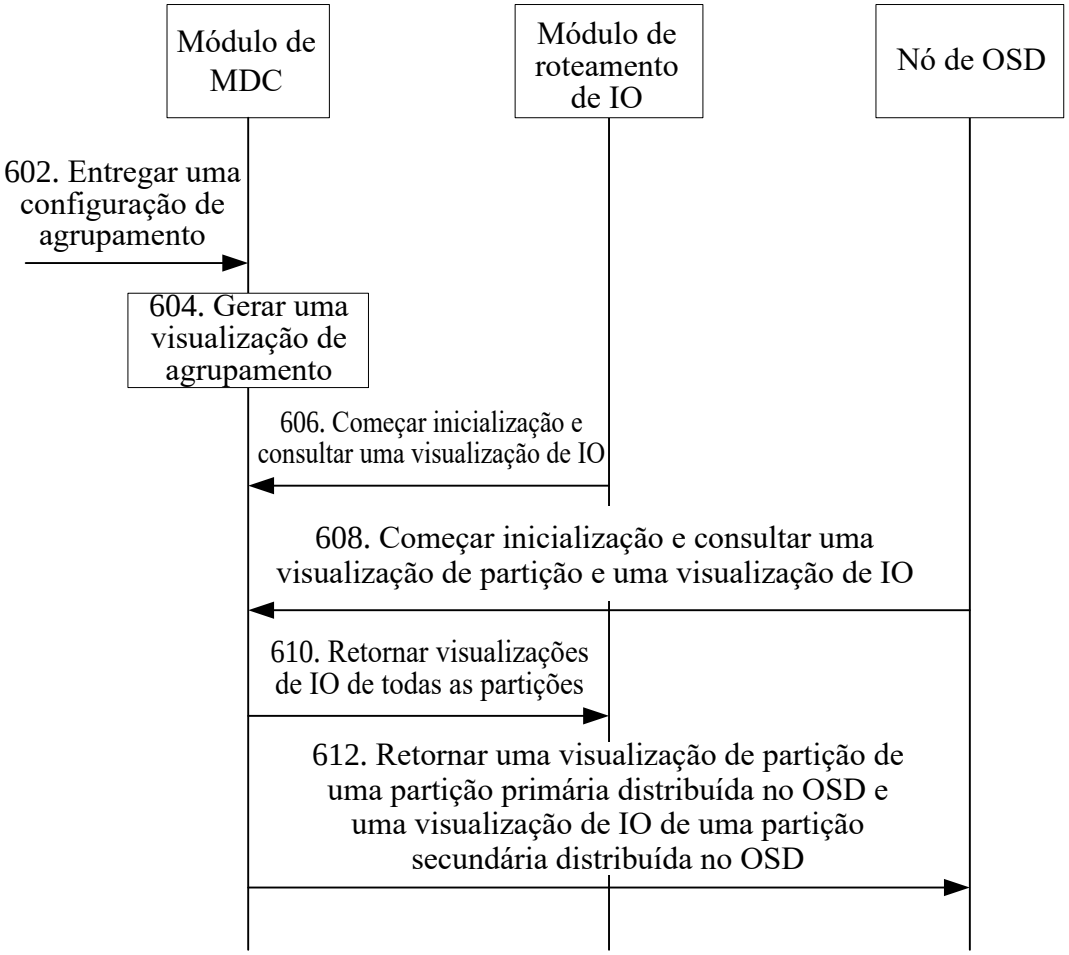

Módulo de MDC
Módulo de roteamento de IO
Nó de OSD
602. Entregar uma configuração de agrupamento
604. Gerar uma visualização de agrupamento
606. Começar inicialização e consultar uma visualização de IO
608. Começar inicialização e consultar uma visualização de partição e uma visualização de IO
610. Retornar visualizações de IO de todas as partições
612. Retornar uma visualização de partição de uma partição primária distribuída no OSD e uma visualização de IO de uma partição secundária distribuída no OSD

FIG. 6

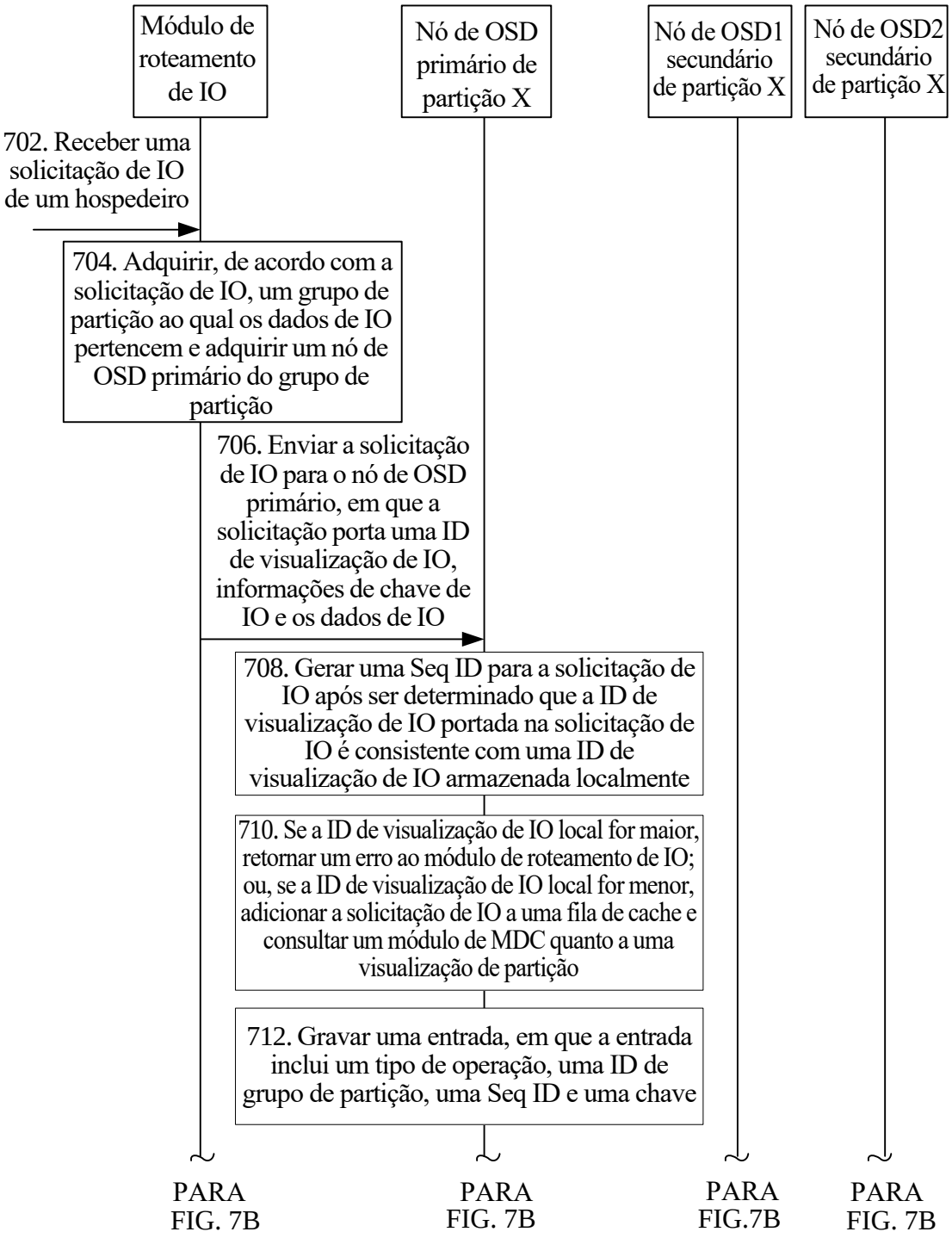

Módulo de roteamento de IO
Nó de OSD primário de partição X
Nó de OSD1 secundário de partição X
Nó de OSD2 secundário de partição X
702. Receber uma solicitação de IO de um hospedeiro
704. Adquirir, de acordo com a solicitação de IO, um grupo de partição ao qual os dados de IO pertencem e adquirir um nó de OSD primário do grupo de partição
706. Enviar a solicitação de IO para o nó de OSD primário, em que a solicitação porta uma ID de visualização de IO, informações de chave de IO e os dados de IO
708. Gerar uma Seq ID para a solicitação de IO após ser determinado que a ID de visualização de IO portada na solicitação de IO é consistente com uma ID de visualização de IO armazenada localmente
710. Se a ID de visualização de IO local for maior, retornar um erro ao módulo de roteamento de IO; ou, se a ID de visualização de IO local for menor, adicionar a solicitação de IO a uma fila de cache e consultar um módulo de MDC quanto a uma visualização de partição
712. Gravar uma entrada, em que a entrada inclui um tipo de operação, uma ID de grupo de partição, uma Seq ID e uma chave
PARA FIG. 7B
PARA FIG. 7B
PARA FIG.7B
PARA FIG. 7B

FIG. 7 A

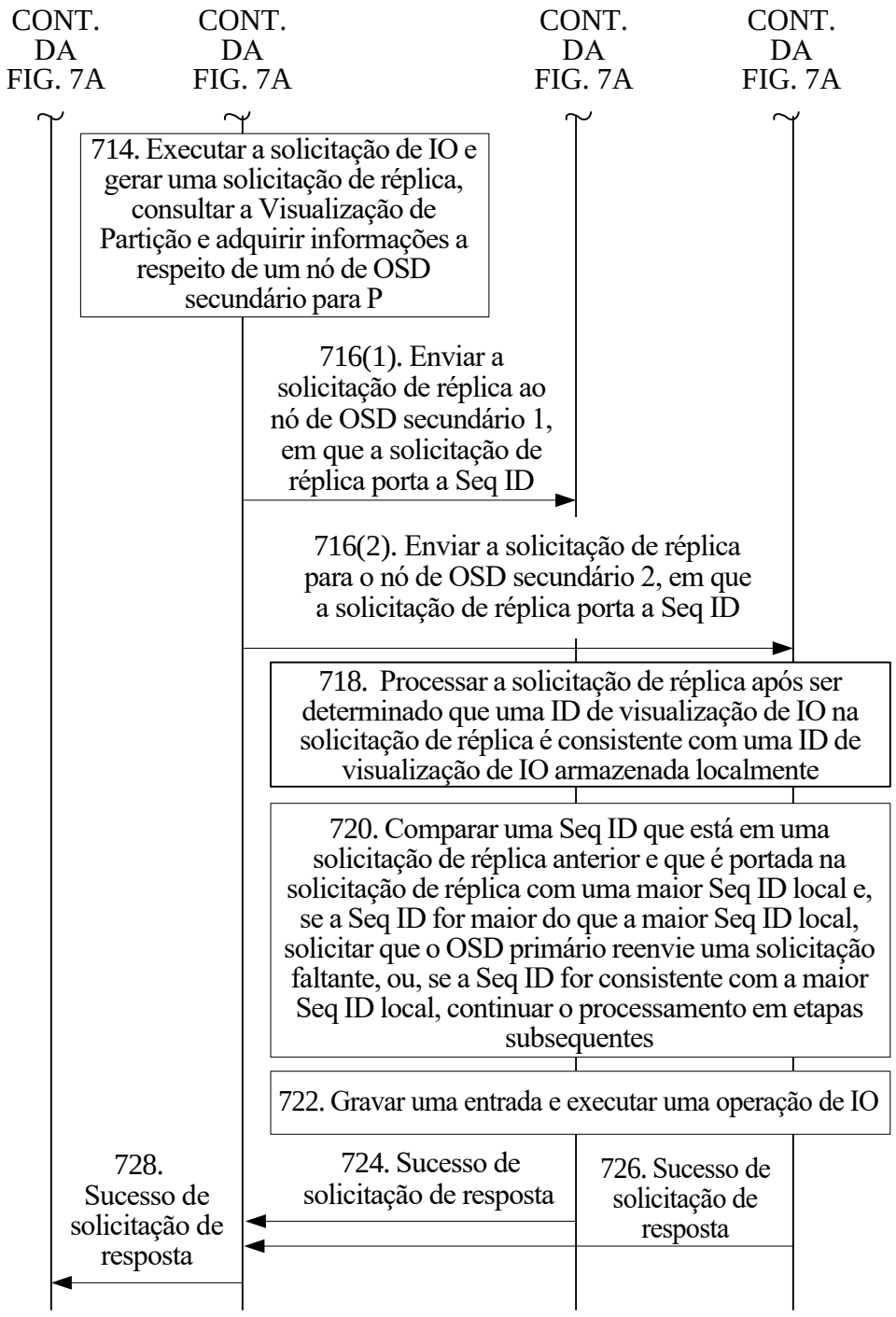

CONT. DA FIG. 7A
CONT. DA FIG. 7A
CONT. DA FIG. 7A
CONT. DA FIG. 7A
714. Executar a solicitação de IO e gerar uma solicitação de réplica, consultar a Visualização de Partição e adquirir informações a respeito de um nó de OSD secundário para P
716(1). Enviar a solicitação de réplica ao nó de OSD secundário 1, em que a solicitação de réplica porta a Seq ID
716(2). Enviar a solicitação de réplica para o nó de OSD secundário 2, em que a solicitação de réplica porta a Seq ID
718. Processar a solicitação de réplica após ser determinado que uma ID de visualização de IO na solicitação de réplica é consistente com uma ID de visualização de IO armazenada localmente
720. Comparar uma Seq ID que está em uma solicitação de réplica anterior e que é portada na solicitação de réplica com uma maior Seq ID local e, se a Seq ID for maior do que a maior Seq ID local, solicitar que o OSD primário reenvie uma solicitação faltante, ou, se a Seq ID for consistente com a maior Seq ID local, continuar o processamento em etapas subsequentes
722. Gravar uma entrada e executar uma operação de IO
728. Sucesso de solicitação de resposta
724. Sucesso de solicitação de resposta
726. Sucesso de solicitação de resposta

FIG. 7 B

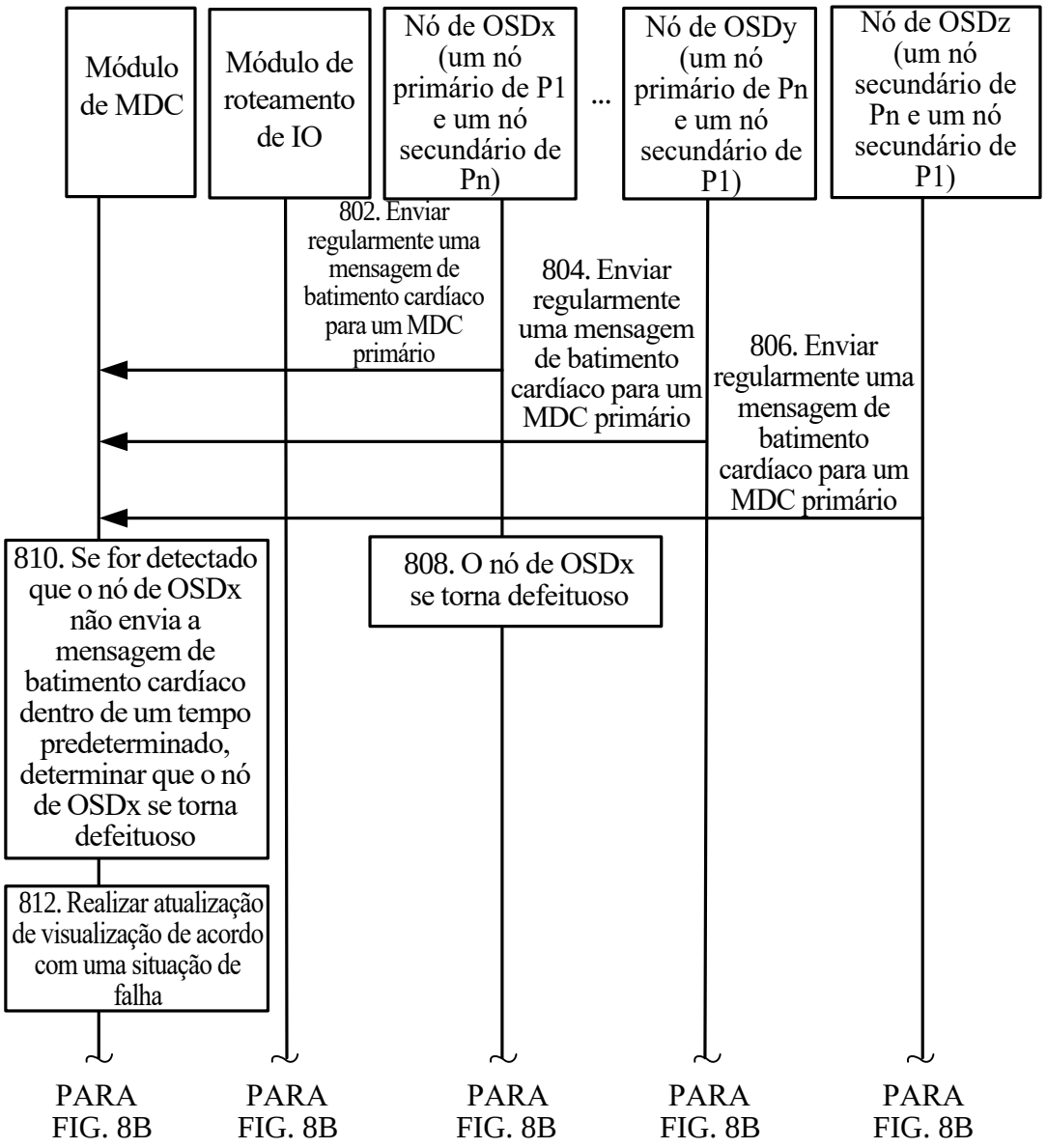

Módulo de MDC
Módulo de roteamento de IO
Nó de OSDx (um nó primário de P1 e um nó secundário de Pn)
...
Nó de OSDy (um nó primário de Pn e um nó secundário de P1)
Nó de OSDz (um nó secundário de Pn e um nó secundário de P1)
802. Enviar regularmente uma mensagem de batimento cardíaco para um MDC primário
804. Enviar regularmente uma mensagem de batimento cardíaco para um MDC primário
806. Enviar regularmente uma mensagem de batimento cardíaco para um MDC primário
808. O nó de OSDx se torna defeituoso
810. Se for detectado que o nó de OSDx não envia a mensagem de batimento cardíaco dentro de um tempo predeterminado, determinar que o nó de OSDx se torna defeituoso
812. Realizar atualização de visualização de acordo com uma situação de falha
PARA FIG. 8B
PARA FIG. 8B
PARA FIG. 8B
PARA FIG. 8B
PARA FIG. 8B

FIG. 8 A

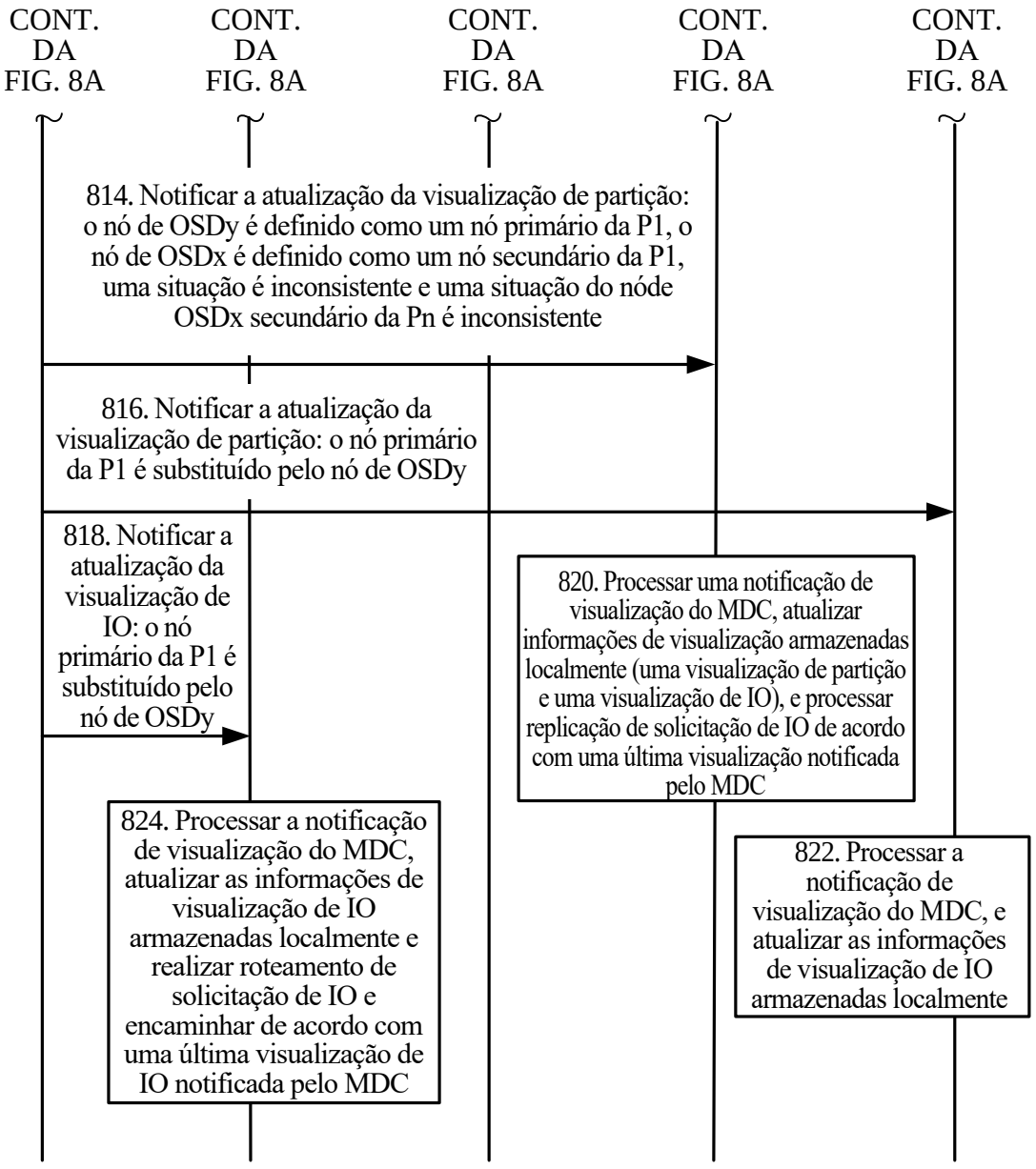

CONT. DA FIG. 8A
CONT. DA FIG. 8A
CONT. DA FIG. 8A
CONT. DA FIG. 8A
CONT. DA FIG. 8A
814. Notificar a atualização da visualização de partição: o nó de OSDy é definido como um nó primário da P1, o nó de OSDx é definido como um nó secundário da P1, uma situação é inconsistente e uma situação do nóde OSDx secundário da Pn é inconsistente
816. Notificar a atualização da visualização de partição: o nó primário da P1 é substituído pelo nó de OSDy
818. Notificar a atualização da visualização de IO: o nó primário da P1 é substituído pelo nó de OSDy
820. Processar uma notificação de visualização do MDC, atualizar informações de visualização armazenadas localmente (uma visualização de partição e uma visualização de IO), e processar replicação de solicitação de IO de acordo com uma última visualização notificada pelo MDC
822. Processar a notificação de visualização do MDC, e atualizar as informações de visualização de IO armazenadas localmente
824. Processar a notificação de visualização do MDC, atualizar as informações de visualização de IO armazenadas localmente e realizar roteamento de solicitação de IO e encaminhar de acordo com uma última visualização de IO notificada pelo MDC

FIG. 8 B

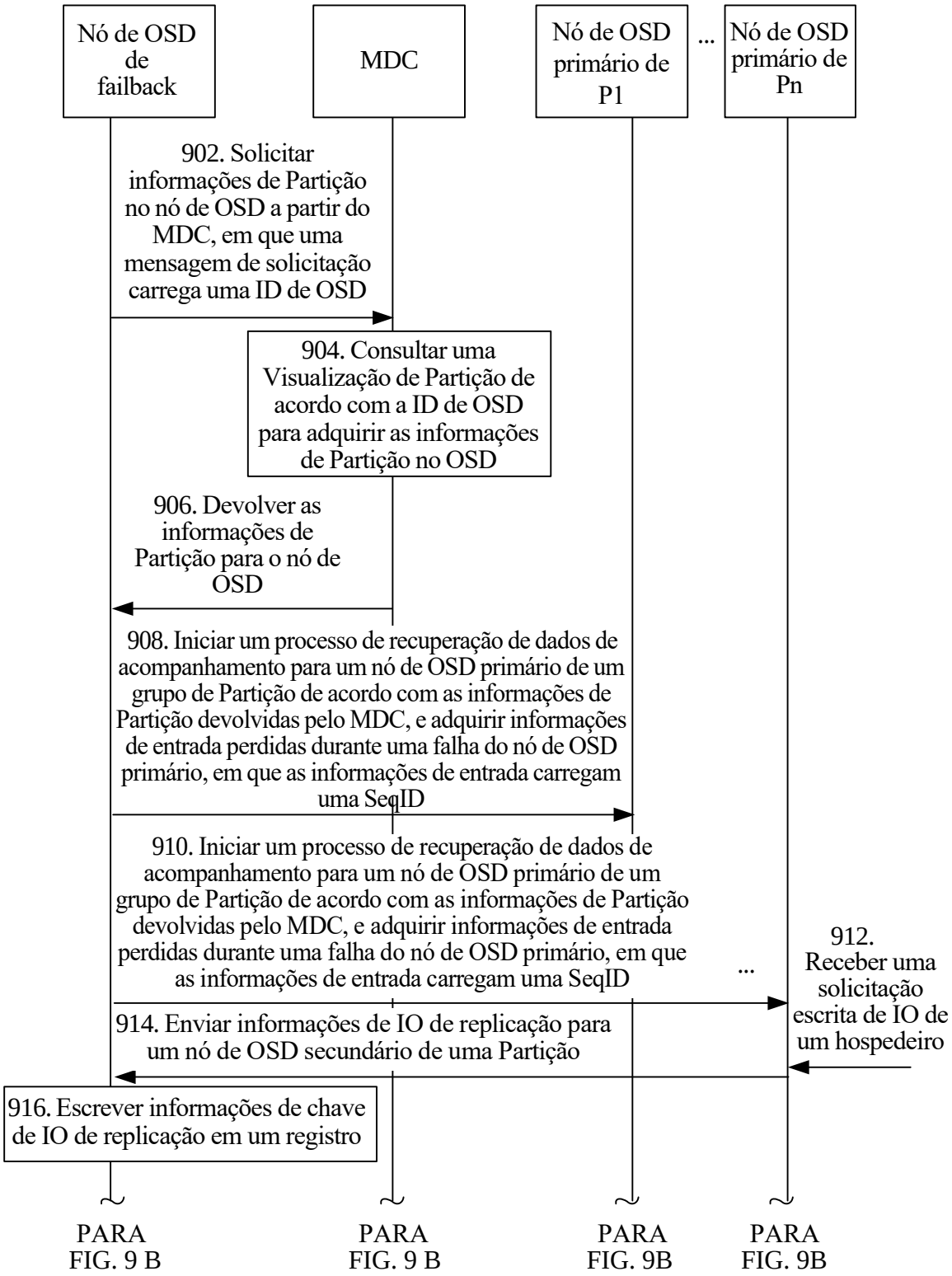

Nó de OSD de failback
MDC
Nó de OSD primário de P1
...
Nó de OSD primário de Pn
902. Solicitar informações de Partição no nó de OSD a partir do MDC, em que uma mensagem de solicitação carrega uma ID de OSD
904. Consultar uma Visualização de Partição de acordo com a ID de OSD para adquirir as informações de Partição no OSD
906. Devolver as informações de Partição para o nó de OSD
908. Iniciar um processo de recuperação de dados de acompanhamento para um nó de OSD primário de um grupo de Partição de acordo com as informações de Partição devolvidas pelo MDC, e adquirir informações de entrada perdidas durante uma falha do nó de OSD primário, em que as informações de entrada carregam uma SeqID
910. Iniciar um processo de recuperação de dados de acompanhamento para um nó de OSD primário de um grupo de Partição de acordo com as informações de Partição devolvidas pelo MDC, e adquirir informações de entrada perdidas durante uma falha do nó de OSD primário, em que as informações de entrada carregam uma SeqID
...
912. Receber uma solicitação escrita de IO de um hospedeiro
914. Enviar informações de IO de replicação para um nó de OSD secundário de uma Partição
916. Escrever informações de chave de IO de replicação em um registro
PARA FIG. 9 B
PARA FIG. 9 B
PARA FIG. 9B
PARA FIG. 9B

FIG. 9 A

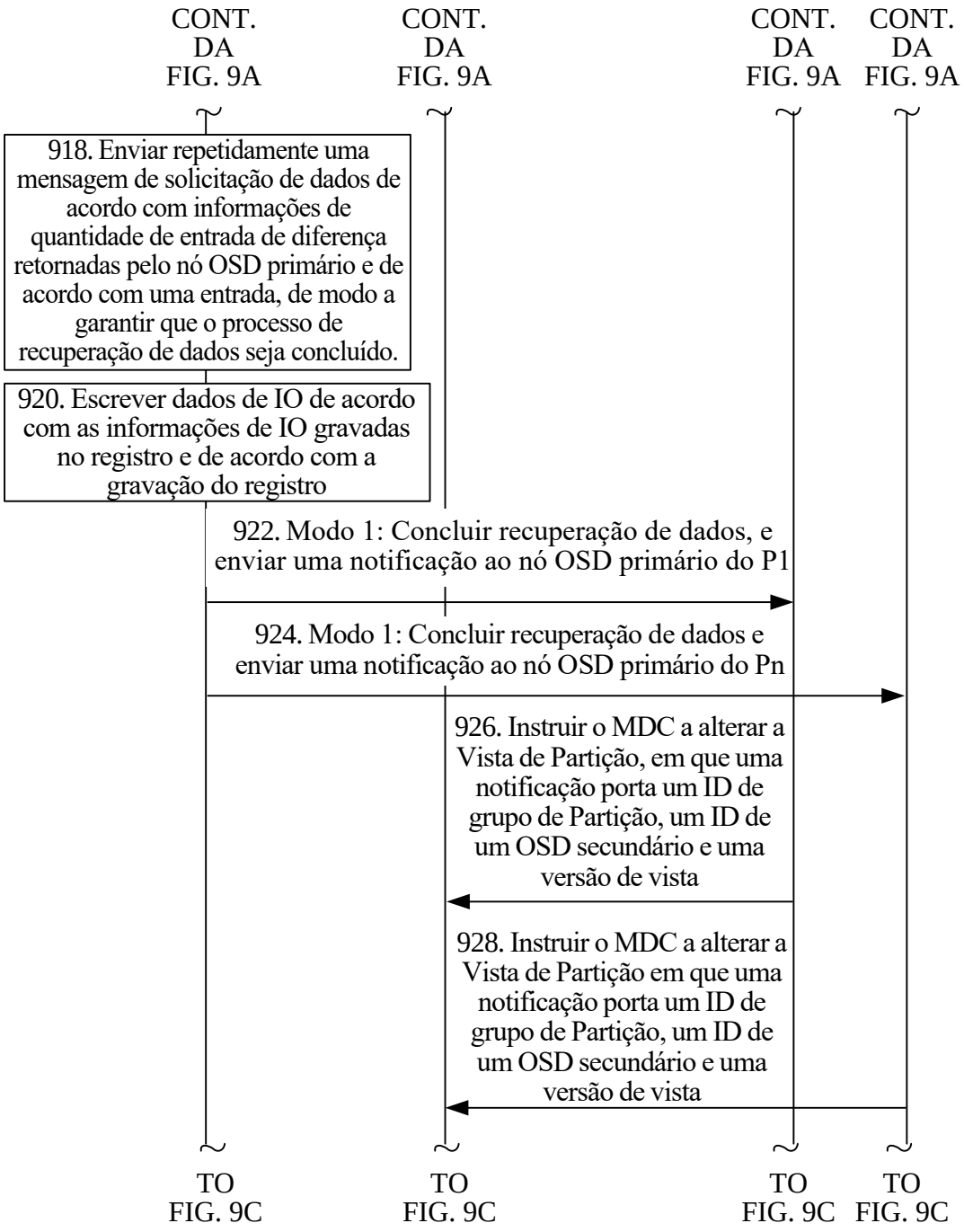

CONT. DA FIG. 9A
CONT. DA FIG. 9A
CONT. DA FIG. 9A
CONT. DA FIG. 9A
918. Enviar repetidamente uma mensagem de solicitação de dados de acordo com informações de quantidade de entrada de diferença retornadas pelo nó OSD primário e de acordo com uma entrada, de modo a garantir que o processo de recuperação de dados seja concluído.
920. Escrever dados de IO de acordo com as informações de IO gravadas no registro e de acordo com a gravação do registro
922. Modo 1: Concluir recuperação de dados, e enviar uma notificação ao nó OSD primário do P1
924. Modo 1: Concluir recuperação de dados e enviar uma notificação ao nó OSD primário do Pn
926. Instruir o MDC a alterar a Vista de Partição, em que uma notificação porta um ID de grupo de Partição, um ID de um OSD secundário e uma versão de vista
928. Instruir o MDC a alterar a Vista de Partição em que uma notificação porta um ID de grupo de Partição, um ID de um OSD secundário e uma versão de vista
TO FIG. 9C
TO FIG. 9C
TO FIG. 9C
TO FIG. 9C

FIG. 9 B

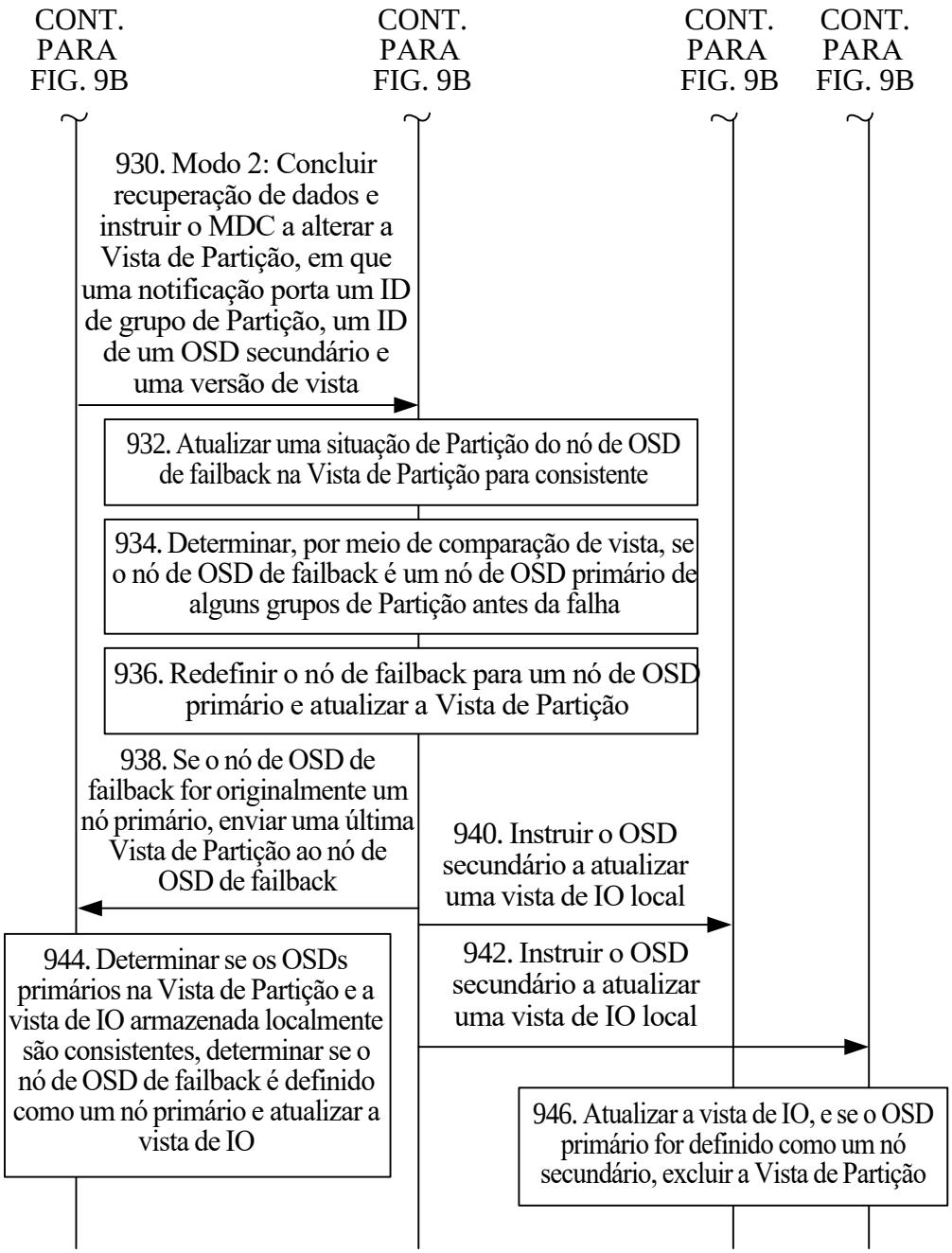

CONT. PARA FIG. 9B
CONT. PARA FIG. 9B
CONT. PARA FIG. 9B
CONT. PARA FIG. 9B
930. Modo 2: Concluir recuperação de dados e instruir o MDC a alterar a Vista de Partição, em que uma notificação porta um ID de grupo de Partição, um ID de um OSD secundário e uma versão de vista
932. Atualizar uma situação de Partição do nó de OSD de failback na Vista de Partição para consistente
934. Determinar, por meio de comparação de vista, se o nó de OSD de failback é um nó de OSD primário de alguns grupos de Partição antes da falha
936. Redefinir o nó de failback para um nó de OSD primário e atualizar a Vista de Partição
938. Se o nó de OSD de failback for originalmente um nó primário, enviar uma última Vista de Partição ao nó de OSD de failback
940. Instruir o OSD secundário a atualizar uma vista de IO local
942. Instruir o OSD secundário a atualizar uma vista de IO local
944. Determinar se os OSDs primários na Vista de Partição e a vista de IO armazenada localmente são consistentes, determinar se o nó de OSD de failback é definido como um nó primário e atualizar a vista de IO
946. Atualizar a vista de IO, e se o OSD primário for definido como um nó secundário, excluir a Vista de Partição

FIG. 9 C

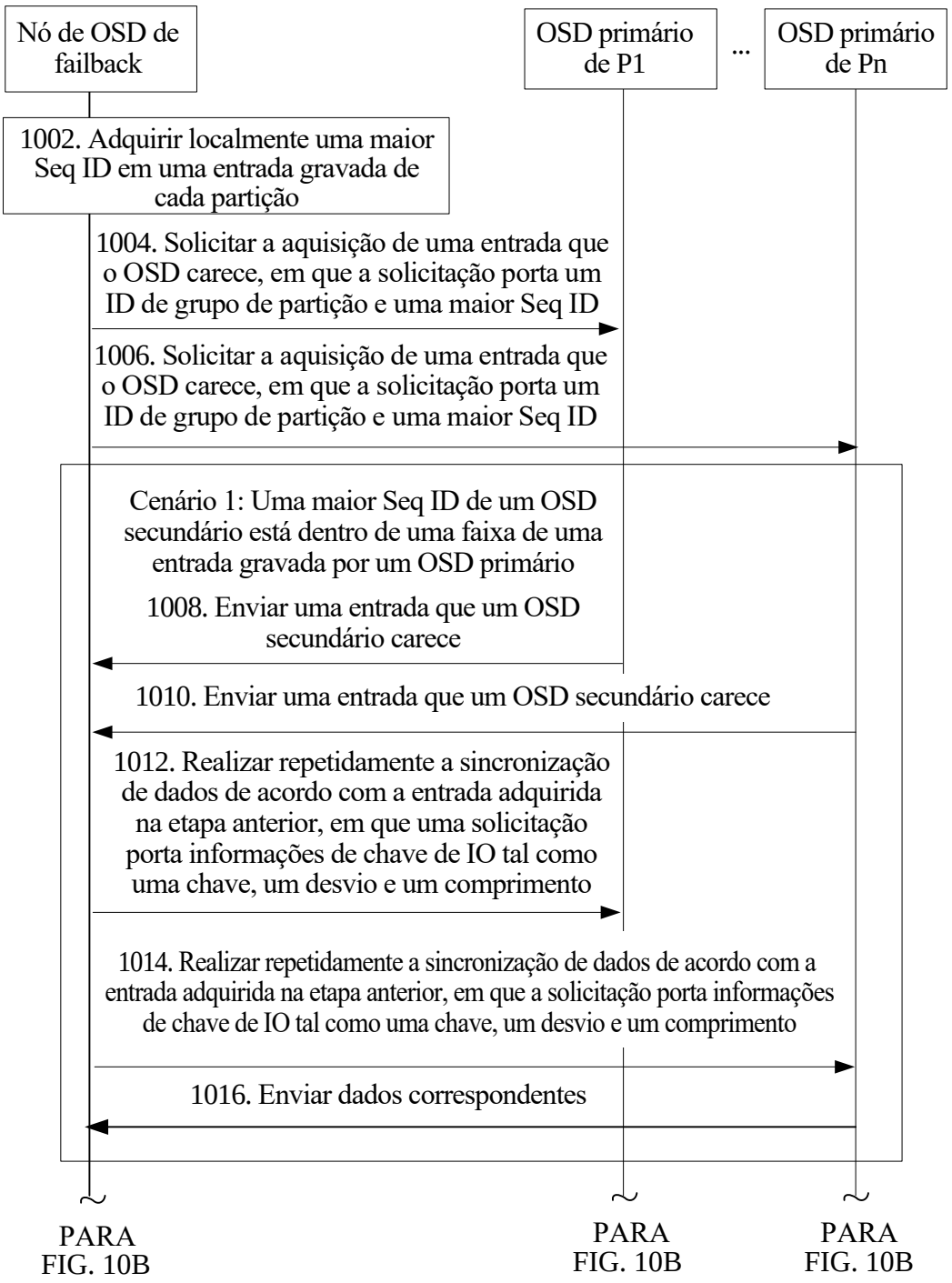

Nó de OSD de failback
OSD primário de P1
...
OSD primário de Pn
1002. Adquirir localmente uma maior Seq ID em uma entrada gravada de cada partição
1004. Solicitar a aquisição de uma entrada que o OSD carece, em que a solicitação porta um ID de grupo de partição e uma maior Seq ID
1006. Solicitar a aquisição de uma entrada que o OSD carece, em que a solicitação porta um ID de grupo de partição e uma maior Seq ID
Cenário 1: Uma maior Seq ID de um OSD secundário está dentro de uma faixa de uma entrada gravada por um OSD primário
1008. Enviar uma entrada que um OSD secundário carece
1010. Enviar uma entrada que um OSD secundário carece
1012. Realizar repetidamente a sincronização de dados de acordo com a entrada adquirida na etapa anterior, em que uma solicitação porta informações de chave de IO tal como uma chave, um desvio e um comprimento
1014. Realizar repetidamente a sincronização de dados de acordo com a entrada adquirida na etapa anterior, em que a solicitação porta informações de chave de IO tal como uma chave, um desvio e um comprimento
1016. Enviar dados correspondentes
PARA FIG. 10B
PARA FIG. 10B
PARA FIG. 10B

FIG. 10 A

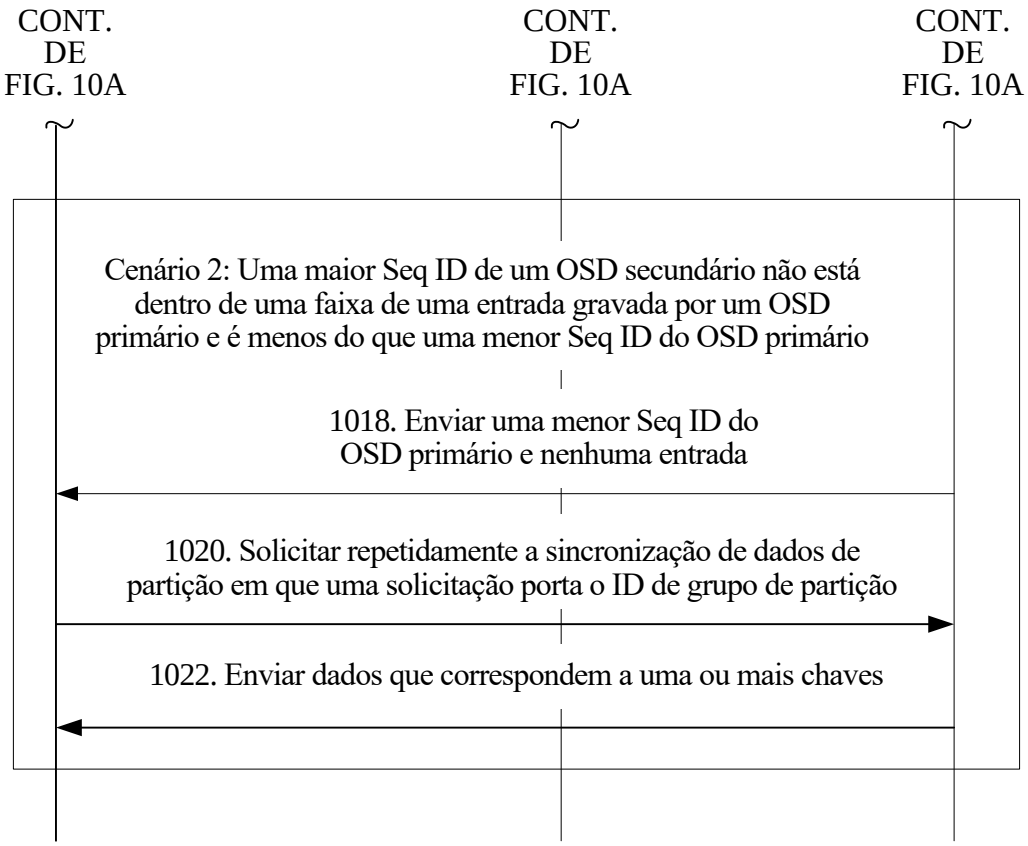

CONT.
DE
FIG. 10A
CONT.
DE
FIG. 10A
CONT.
DE
FIG. 10A
Cenário 2: Uma maior Seq ID de um OSD secundário não está dentro de uma faixa de uma entrada gravada por um OSD primário e é menos do que uma menor Seq ID do OSD primário
1018. Enviar uma menor Seq ID do OSD primário e nenhuma entrada
1020. Solicitar repetidamente a sincronização de dados de partição em que uma solicitação porta o ID de grupo de partição
1022. Enviar dados que correspondem a uma ou mais chaves

FIG. 10 B

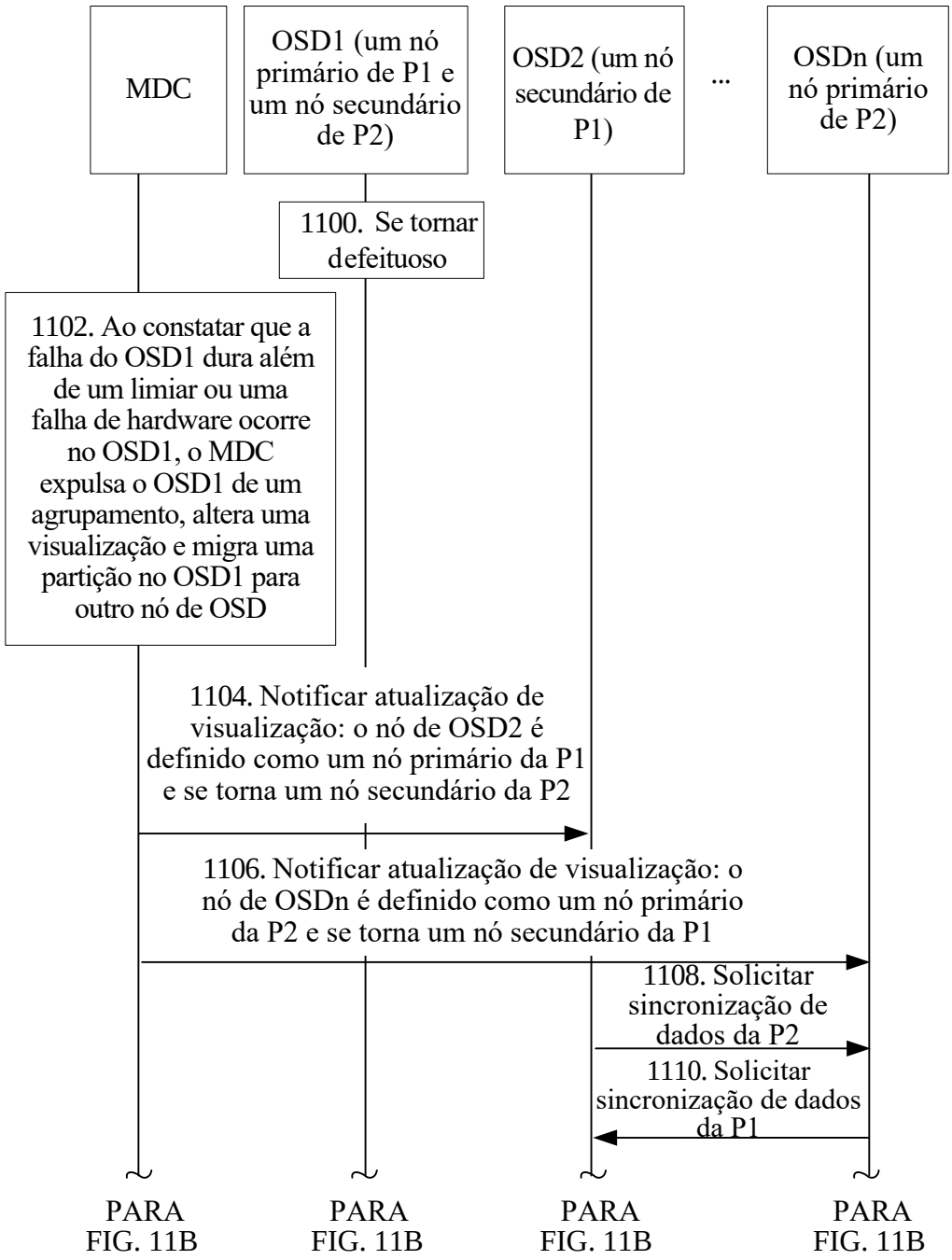

MDC
OSD1 (um nó primário de P1 e um nó secundário de P2)
OSD2 (um nó secundário de P1)
...
OSDn (um nó primário de P2)
1100. Se tornar defeituoso
1102. Ao constatar que a falha do OSD1 dura além de um limiar ou uma falha de hardware ocorre no OSD1, o MDC expulsa o OSD1 de um agrupamento, altera uma visualização e migra uma partição no OSD1 para outro nó de OSD
1104. Notificar atualização de visualização: o nó de OSD2 é definido como um nó primário da P1 e se torna um nó secundário da P2
1106. Notificar atualização de visualização: o nó de OSDn é definido como um nó primário da P2 e se torna um nó secundário da P1
1108. Solicitar sincronização de dados da P2
1110. Solicitar sincronização de dados da P1
PARA FIG. 11B
PARA FIG. 11B
PARA FIG. 11B
PARA FIG. 11B

FIG. 11 A

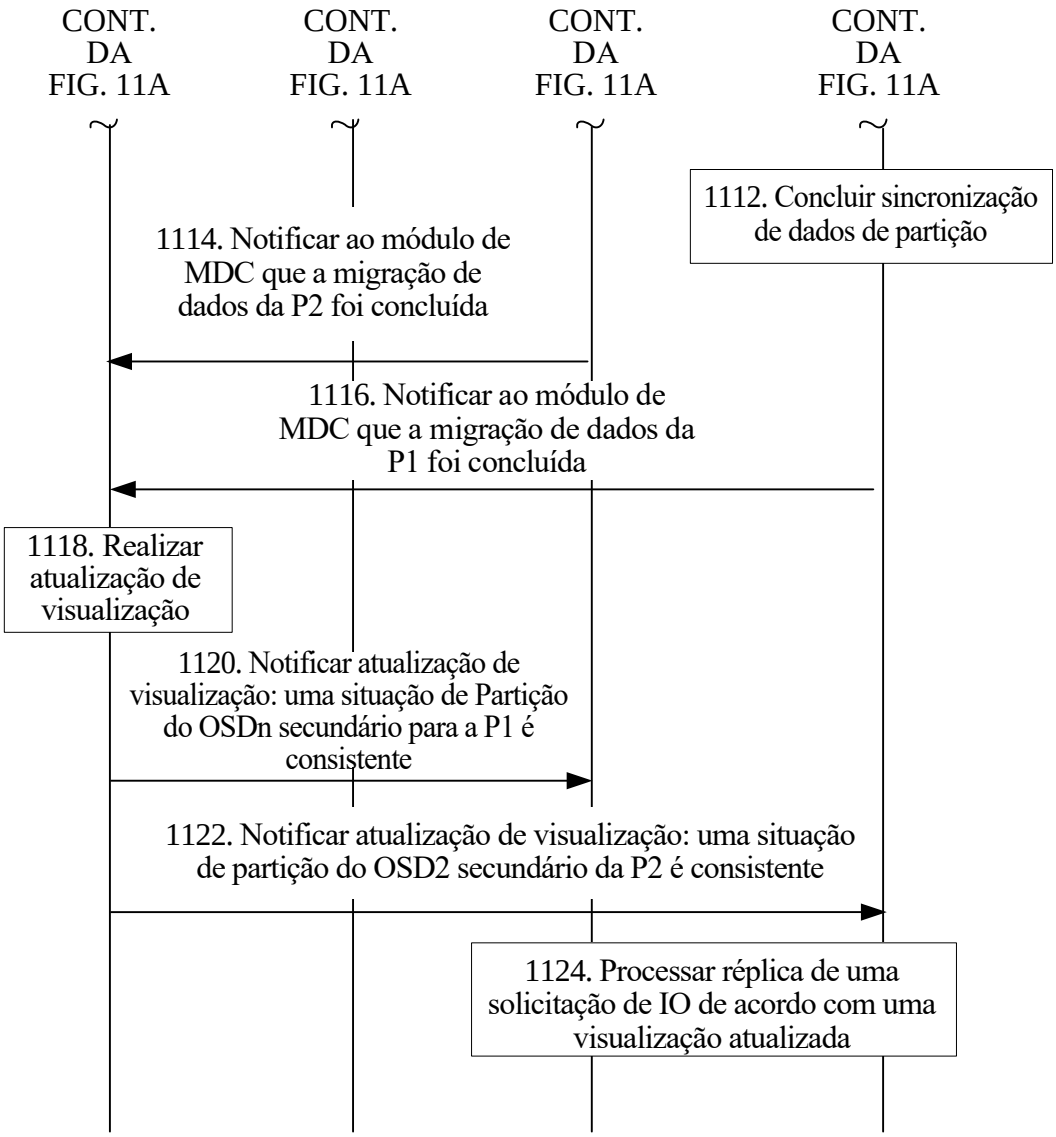

CONT. DA FIG. 11A
CONT. DA FIG. 11A
CONT. DA FIG. 11A
CONT. DA FIG. 11A
1112. Concluir sincronização de dados de partição
1114. Notificar ao módulo de MDC que a migração de dados da P2 foi concluída
1116. Notificar ao módulo de MDC que a migração de dados da P1 foi concluída
1118. Realizar atualização de visualização
1120. Notificar atualização de visualização: uma situação de Partição do OSDn secundário para a P1 é consistente
1122. Notificar atualização de visualização: uma situação de partição do OSD2 secundário da P2 é consistente
1124. Processar réplica de uma solicitação de IO de acordo com uma visualização atualizada

FIG. 11 B

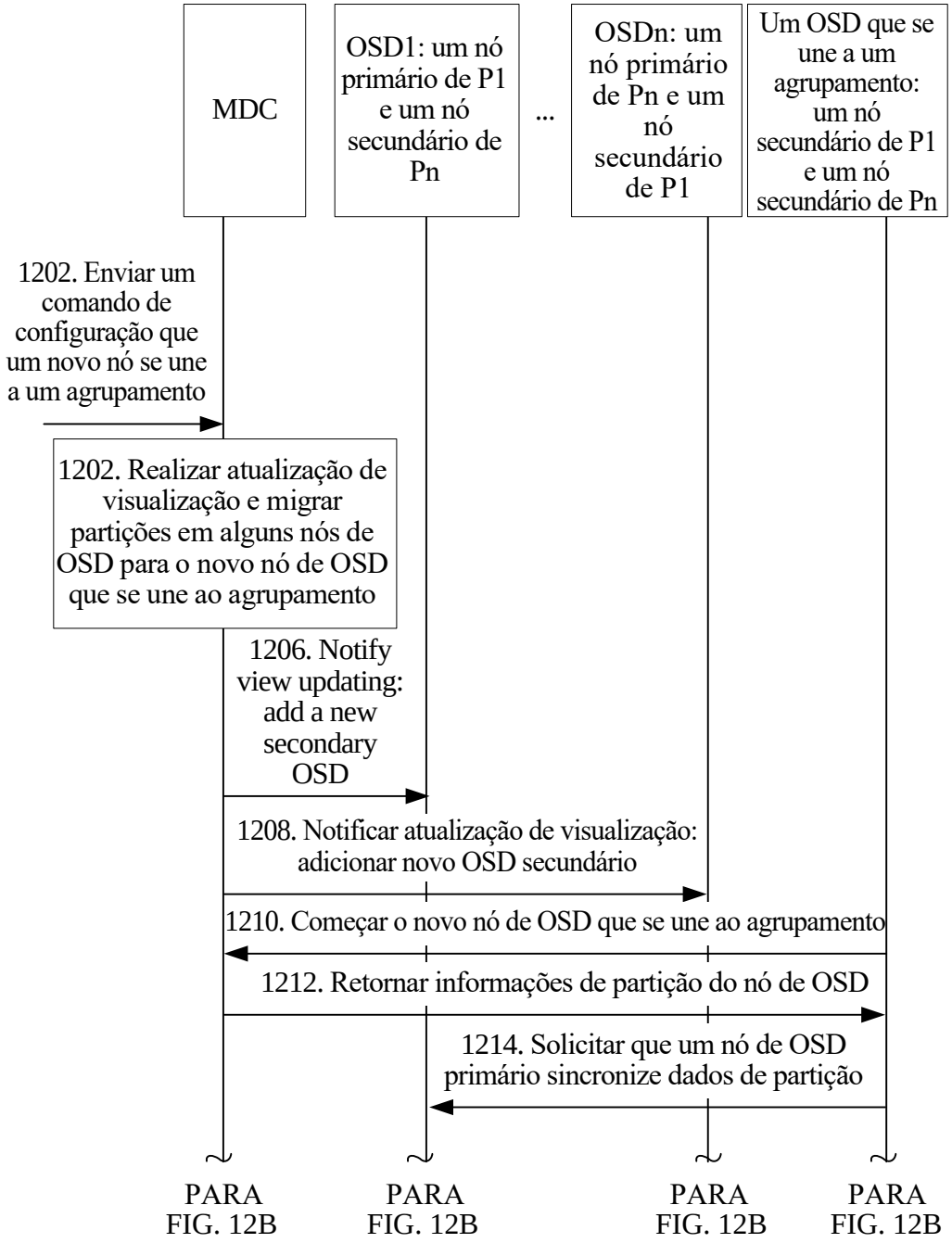

MDC
OSD1: um nó primário de P1 e um nó secundário de Pn
...
OSDn: um nó primário de Pn e um nó secundário de P1
Um OSD que se une a um agrupamento: um nó secundário de P1 e um nó secundário de Pn
1202. Enviar um comando de configuração que um novo nó se une a um agrupamento
1202. Realizar atualização de visualização e migrar partições em alguns nós de OSD para o novo nó de OSD que se une ao agrupamento
1206. Notify view updating: add a new secondary OSD
1208. Notificar atualização de visualização: adicionar novo OSD secundário
1210. Começar o novo nó de OSD que se une ao agrupamento
1212. Retornar informações de partição do nó de OSD
1214. Solicitar que um nó de OSD primário sincronize dados de partição
PARA FIG. 12B
PARA FIG. 12B
PARA FIG. 12B
PARA FIG. 12B

FIG. 12 A

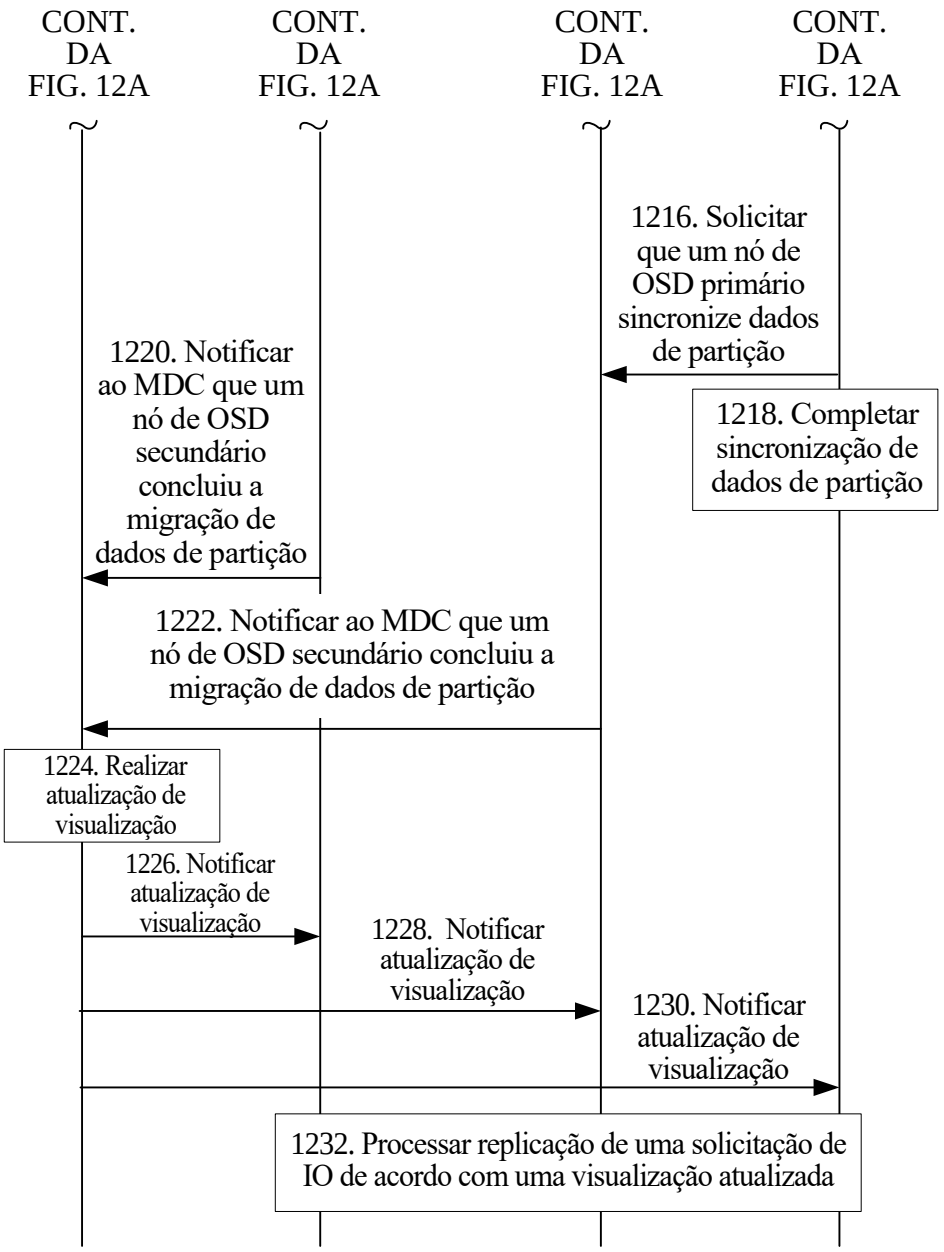

CONT. DA FIG. 12A
CONT. DA FIG. 12A
CONT. DA FIG. 12A
CONT. DA FIG. 12A
1216. Solicitar que um nó de OSD primário sincronize dados de partição
1218. Completar sincronização de dados de partição
1220. Notificar ao MDC que um nó de OSD secundário concluiu a migração de dados de partição
1222. Notificar ao MDC que um nó de OSD secundário concluiu a migração de dados de partição
1224. Realizar atualização de visualização
1226. Notificar atualização de visualização
1228. Notificar atualização de visualização
1230. Notificar atualização de visualização
1232. Processar replicação de uma solicitação de IO de acordo com uma visualização atualizada

FIG. 12 B