MEMORIAL DESCRITIVO

1. OBJETO

Aquisição de equipamentos de infraestrutura de *storage* para dados científicos obtidos e derivados das linhas de luz do projeto Sirius. A proposta deve considerar a dinâmica de geração e uso de dados científicos para projetos de pesquisa e desenvolvimento apoiando usuários internos e externos associados às linhas de luz do Sirius.

2. INTRODUÇÃO

A empresa deverá elaborar cenários que atendam aos requisitos mínimos descritos neste documento, *evidenciando* as funcionalidades desejadas com o desempenho esperado para o projeto Sirius.

3. GENERALIDADES

O crescente uso de dados nas instalações do Laboratório Nacional de Luz Síncrotron (LNLS) cria a demanda por armazenamento compartilhado entre suas estações experimentais (chamadas linhas de luz) e demais subprojetos associados do Sirius. Dentro da missão do centro, a manipulação de dados por usuários internos e externos também consome espaço para esta unidade compartilhada. Na competição pelo uso do armazenamento, há uma diferença entre os tipos de usuários que acessam dados. Está previsto um cenário em que o armazenamento de alto desempenho está localizado no data center (DC), dentro do edifício do Sirius. Este data center se conecta às linhas de luz por meio de uma fibra 100 Gbps ethernet. O data center é composto de 3 ilhas, e cada ilha possui 12 racks, divididos entre nós de processamento computacional, armazenamento e rede, todos conectados entre si for fibra ethernet de 100 Gbps. O impacto direto no *storage* ocorre devido aos seguintes grupos: linha de luz, pesquisadores, usuários internos e usuários externos. Cada um destes grupos gera dados de acordo com uma taxa por unidade de tempo, que pode ser estimada a priori por um limite superior de acordo com o histórico de uso da instalação e com a expectativa de uso do Sirius. A maioria destes grupos interage com algum tipo de máquina de alto desempenho, gerando dados a partir de medições em instalações internas ao LNLS. O uso de acordo com cada grupo é dado por:

1. *Linha de luz*: existe uma taxa de geração diária de Terabytes, que é enviada para armazenamento em duas situações, já processadas pela máquina de alto desempenho localizada na linha de luz ou a ser processada pelo cluster no DC.
2. *Apoio e Pesquisa*: Algumas instalações internas podem gerar dados com volumetria semelhante às linha de luz. Por exemplo, dados que já estão armazenados e serão processados posteriormente (seguindo alguma fila de processamento) no cluster aumentam significativamente o consumo de espaço do *storage*;

3. *Usuários*: Dois tipos de usuários são identificados, interno e externo. Ambos utilizam dados já coletados e armazenados para processamento computacional, gerando mais dados que consomem a capacidade total do *storage*.

Para cada grupo acima, há três ações que movem o fluxo de dados no armazenamento, conforme ilustrado na Figura 1; são estes (a) *upload* de dados no *storage* (b) processamento de dados contidos no *storage*, e (c) *download* de dados armazenados no *storage*. Estas três ações ocorrem de forma aleatória dentro da dinâmica do projeto Sirius. Idealmente, as taxas de leitura e gravação variam para cada ação (a), (b) e (c); por exemplo, a taxa de gravação é maior que a taxa de leitura na ação (a), enquanto o oposto deve ser verdadeiro para (b). De acordo com a experiência e o histórico do LNLS, entende-se que a porcentagem atribuída a cada ação para cada grupo de uso é conforme mostrada na Tabela 1. Aqui, assumimos que as linhas de luz são divididas igualmente entre upload e processamento de cluster via agendador de tarefas, diretamente impactando o armazenamento. Isso já não é verdade para grupos de usuários que não fazem *upload*, mas fazem o *download* de dados processados. Cada grupo tem um número máximo de pessoas e servidores; assim como cada servidor usa um tipo diferente de protocolo para acesso ao armazenamento. Os sistemas operacionais normalmente usados pelos grupos para acesso a dados são Unix e Windows, o que afeta o sistema de acesso ao armazenamento, que deve ser fácil e modular. Conforme apresentado na Tabela 1, assumimos um limitante superior de 100 usuários por dia acessando o *storage*, de acordo com cada grupo; gerando uma quantidade de acesso global de pelo menos 1000 usuários por dia, dentro da dinâmica apresentada na Figura 1.

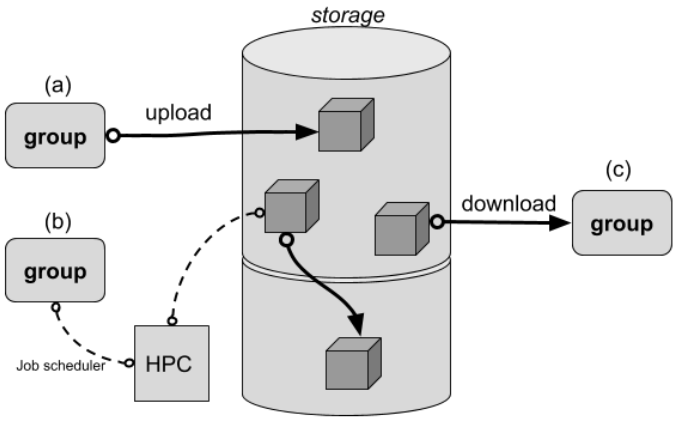

Figura 1: Três ações diferentes para um determinado grupo acessando o armazenamento: download, upload e através do sistema HPC via *job scheduler*.

Os requisitos para o projeto e a especificação dos dados de rede para as linhas de luz são fundamentados nos detectores desenvolvidos pelo LNLS. Cada detector terá pelo menos uma interface de rede de 100 Gbps conectada diretamente a um servidor de alto desempenho localizado fisicamente na linha de luz. Após o processamento neste servidor, os dados processados

Grupo		Linhas de Luz							Staff	Usuarios	
		CAT	MAN	MOG	CAR	EMA	IPE	BL		I	E
Ação (%)	(a)	50	50	50	50	50	50	50	40	10	10
	(b)	50	50	50	50	50	50	50	40	80	50
	(c)	0	0	0	0	0	0	0	20	10	40
Acessos/dia		100	100	100	100	100	100	100	100	100	100

Tabela 1: Portagem de ação para alguns dos principais grupos de acesso no storage global

precisam ser carregados em um armazenamento central que suporte uma alta taxa de transferência; sua localização afeta diretamente o design da rede, para que o posicionamento do *storage* contemple toda a vazão de dados de projetos científicos. A maior parte do processamento de dados do LNLS é baseada em GPU devido ao ganho de eficiência computacional, juntamente com o baixo consumo energético. Os nós do servidor de processamento local suportam um rápido processamento de uma tal forma que um sistema de armazenamento rápido (tipicamente paralelo) também será crucial para o fluxo adequado dos dados medidos. Após o processamento dos dados nos servidores locais, os dados experimentais precisam ser transferidos da memória para o sistema de armazenamento, para que a linha de luz esteja apta para receber novos dados experimentais. Por exemplo, um detector de área com n x n pixels obtendo a=1536 projeções em um experimento de imagens gera um conjunto de dados cúbicos de 13GB com precisão de ponto flutuante (1536^3 voxels) por amostra por alguns segundos. Este tempo de processamento rápido é possível devido à arquitetura multi GPU do servidor local. Durante um conjunto destes experimentos, um servidor local com 1 TB de memória pode armazenar dados de mais de 50 medidas experimentais antes de precisar descarregá-las no *storage* principal a uma alta taxa de escrita. Para aplicações mais exigentes em processamento produzindo medidas de 3072^3 voxels por unidades de minuto, cerca de 0,5 TB de dados são gerados, podendo alcançar 5TB por dia. Nestes casos, o upload de dados para armazenamento será feito de forma sequencial para o *storage*. Considerando a notação da Tabela 2 e duas linhas de luz de imagem com detectores com n=1536 e n=3072 pixels, os dados medidos chegam ao detector com 12 ou 24 bits e são convertidos para 32 bits, conforme indicado na etapa (1). O servidor local precisará armazenar na memória uma quantidade de dados que depende claramente do número de ângulos obtidos (projeções, indicado por a). Mais ângulos têm o benefício de proporcionar uma reconstrução fiel usando determinadas estratégias algorítmicas, ao passo que exigem mais memória. Dois exemplos de medição são apresentados em que o servidor local deverá conter 47 ou 378 Gigabytes de dados no melhor caso, respectivamente. Supondo que uma escolha de profundidade de bits seja definida anteriormente para reconstrução e segmentação, os

Throughput per experiment

n			1536											
Step			[1]	[2]					[3]		[4]			
Datasets			Measure (GB)	Reconstruction (GB)			Segment. (GB)		Throughput (GB/s)		Flow (sec)			
			Bit depth						Lower bound	Upper bound	Lower Bound		Upper bound	
			32	8	16	32	32	8			100 Gbs	70 Gbs	100 Gbs	70 Gbs
a = *h n*	*h*	3	40.5	3.375	6.75	13.5	13.5	3.375	47.25	67.5	3.78	5.40	5.40	7.71
		2	27						33.75	54	2.70	3.86	4.32	6.17
		1	13.5						20.25	40.5	1.62	2.31	3.24	4.63
		0.5	6.75						13.5	33.75	1.08	1.54	2.70	3.86
n			3072											
Step			[1]	[2]					[3]		[4]			
Datasets			Measure (GB)	Reconstruction (GB)			Segment. (GB)		Throughput (GB/s)		Flow (sec)			
			Bit depth						Lower bound	Upper bound	Lower Bound		Upper bound	
			32	8	16	32	32	8			100 Gbs	70 Gbs	100 Gbs	70 Gbs
a = *h n*	*h*	3	324	27	54	108	108	27	378	540	30.24	43.2	43.20	61.7
		2	216						216	216	17.28	24.6	17.28	24.6
		1	108						108	108	8.64	12.3	8.64	12.3
		0.5	54						54	54	4.32	6.1	4.32	6.17

Tabela 2: Exemplo de *throughput* por experimento em linhas de imagem.

dados processados também podem variar em tamanho, entre 67,5 e 540 gigabytes para detectores de 1536² e 3072² no pior caso, respectivamente. Conforme indicado na Tabela 2, o fluxo de 67,5 GB de dados (medida e dados processados) deve levar - nos piores casos - 5,4 segundos para desocupar o servidor principal e chegar no *storage* utilizando algum sistema de escrita paralela. Estas taxas dependem fortemente do tipo de hardware utilizado e do sistema de paralelismo desenvolvido pelo sistema global de armazenamento. A solução de *storage* deve aguentar esta vazão de dados, proveniente das várias linhas de luz, de uma forma escalável e modular.

4. REQUISITOS MINIMOS

A proposta de *storage* para atender as demandas do projeto Sirius estão listadas na tabela abaixo. Cada uma das categorias apresentadas possui um peso diferente, de acordo com a prioridade dentro das demandas de uso e processamento de dados das linhas de luz e projetos correlatos. O data center do projeto Sirius já contempla a infraestrutura de Racks, apta a receber uma proposta de solução em *storage*, do qual cada rack possui 47U com espelhamento de cabos de rede (para gerência e fibra, descontando 2U), com 800 x 800 mm, com perfil de 19" ajustável para comportar diferentes tipos de equipamentos. A solução proposta deve conter cabos C13/C14 para energia e deve também incluir Gbic's (multimodo) para comunicação com o switch Core de acordo com a solução proposta e somente o Gbic de conexão no lado da solução de *Storage*.

NOTA: O espaço líquido total disponível da solução de *storage* deve ser híbrida com 80% dos discos mecânicos e 20% flash ou 100% flash.

Categoria	*Requisito*	*Justificativa*
Capacidade	Inicial com 2 Petabytes num espaço de 1 rack (47U)	O centro de processamento de dados do LNLS/Sirius necessita de um sistema inicial de 2 Petabytes (PB) que atenda à demanda inicial das primeiras linhas da fase A do projeto Sirius. A solução de 2PB deve estar alocada em no máximo 1 rack (composto de 47U) já existente na infra-estrutura do Sirius.
	Escalável a 30 Petabytes de capacidade líquida	A solução proposta deve alcançar até 30 PB futuramente, de uma forma elástica, com baixa granularidade.
Performance	File system paralelo para acesso	O sistema de acesso aos dados deve ser paralelo, permitindo rápido acesso aos dados, para cada grupo de uso, conforme descrito na seção 3.

	Gerenciamento de blocos escalável	O gerenciamento global do *storage* deve ser amigável através de uma interface configurável e de fácil manuseio, que permita gerenciar acessos, quotas, permissões, blocos de escrita e leitura
	Acessível por no mínimo 1000 usuários	A solução deve contemplar um mínimo de 1000 conexões acessando globalmente o *storage* principal, em todas as categorias de grupos descritas na Seção 3.
	Taxa de escrita > 200 Gbps	Considerando as principais linhas da fase A do Sirius, cujo máximo throughput pode atingir uma taxa nominal de 5TB por dia, as taxas de escrita devem ser no mínimo de 200 Gbps.
	Taxa de leitura > 100 Gbps	Considerando as principais linhas da fase A do Sirius, cujo máximo throughput pode atingir uma taxa nominal de 5TB por dia, as taxas de leitura devem ser no mínimo de 200 Gbps.
	Tierização nativa para dados quentes e frios	A solução de *storage* deve contemplar um sistema de tierização nativo. Todo o hardware e software necessários para a divisão de dados frios e quentes, de acordo com a política do LNLS/Sirius, devem estar contidos na proposta.
Acesso	CIFS e SMB	A proposta deve contemplar protocolos que sejam capaz de suportar usuários em diferentes sistemas operacionais. Todo hardware excedente alocado para esta finalidade, e que faça parte da solução, deve estar totalmente integrado e desenvolvido para atender este requisito.
	NFS	A maioria dos serviços internos ao LNLS são baseados em sistemas Unix, portanto o sistema deve contemplar este protocolo de uma forma nativa. Todo hardware excedente alocado para esta finalidade, e que faça parte da solução, deve estar totalmente integrado e desenvolvido para

		atender este requisito.
	HTTP	Este protocolo atenderá usuários externos e internos com o acesso aos dados contidos no *storage*.
	Flexível e configurável	As permissões a cada grupo de acesso devem ser facilmente configuráveis pela solução, sem a necessidade de parada do hardware.
Eficiência	Sistema operacional consistente	A consistência da solução deve ser tal, que todas as ferramentas necessárias para seu desempenho estejam contidas, sem a necessidade de hardware e software extra.
	Deduplicação nativa	Aliviar a utilização de dados na *storage*
	Compressão nativa	Isenção de redundância nos arquivos compactados.
	Quotas em tempo real	O provisionamento de quotas para cada um dos grupos de acesso ao *storage* deve ser de fácil manuseio e configuração, sem a necessidade de parada do hardware.
Segurança dos dados	Disaster recovery	Cópia fria dos arquivos da *storage* em produção para uma *storage* reserva
	Disk fail	Garantir a disponibilidade dos arquivos mesmo quando um disco/controladora entra em falha
	Integrity assurance	Garantir a disponibilidade dos arquivos mesmo quando um disco/controladora entra em falha
Escalabilidade	Modular e flexível	Possível upgrade de volumes sem a necessidade de adquirir uma nova solução completa.
Segurança da informação	LDAP	É o protocolo de login utilizado internamente no CNPEM
	Criptografia	Garantia de que dados sejam trafegados na rede de forma criptografada.

	Segurança na interface de gerenciamento	Todo o acesso a gerência da *storage* deve ser criptografada.
	Controle de acessos	Garantia de que somente quem é autorizado possa acessar arquivos na unidade de *storage*.
	Auditoria	Ferramentas de relatórios com informações sobre quem acessou, quando aconteceu e se houve alteração nos dados.
Implementação	Instalação, configuração e treinamento	A solução deve ser capaz de instalar todo o hardware, configurá-lo de acordo com os requisitos do LNLS e prover treinamento aos times responsáveis do CNPEM.
	Suporte 24/7/365 x 4 por 5 anos	Atendimento em caso de *upgrade*, bugs e/ou falhas de hardware.
Politica de dados	Tipos de dados (a) varios arquivos pequenos (b) um único arquivo grande	A solução deve ser capaz de transferir dados com diferentes tipos de tamanhos, desde uma seqüência de arquivos muito pequenos (milhões de arquivos variando entre KB e MB), até únicos arquivos com dezenas de terabytes - ambos a taxas semelhantes.
	Transferência rápida	O sistema de paralelismo deve possuir uma boa performance de transferência para quaisquer tipos de arquivos, não sendo específico para determinadas aplicações.
Conexão	Ethernet	O sistema deve contemplar o uso do protocolo Ethernet. Toda a infraestrutura de conexão somente do lado do storage deve estar contida na solução
	Infiniband	A solução deve ser adaptável a soluções com o protocolo Infiniband para o futuro, mesmo não sendo a escolha inicial do projeto Sirius.

5. PRAZOS

Data máxima para envio da proposta comercial e cenário com detalhamento de equipamentos: 08/02/2020

6. CONTATOS PARA ESCLARECIMENTO DE QUESTÕES TÉCNICAS

Felipe Campos de Oliveira
Especialista em redes
Tecnologia da Informação e Comunicação - TIC
Centro Nacional de Pesquisa em Energia e Materiais (CNPEM)
(19) 99822-7877, (19) 3517.5181
felipe.campos@cnpem.br
www.cnpem.br

Eduardo Xavier Miqueles
Pesquisador
Grupo de Computação Científica Sirius - GCC
Centro Nacional de Pesquisa em Energia e Materiais (CNPEM)
(19) 3512.1043
eduardo.miqueles@lnls.br
www.cnpem.br

James Piton
Especialista
Software de Operação das Linhas - SOL
Centro Nacional de Pesquisa em Energia e Materiais (CNPEM)
(19) 3512.1228
james@lnls.br
www.cnpem.br

Sérgio A. Carrare Jr.
Analista de Infraestrutura
Tecnologia da Informação e Comunicação - TIC
Centro Nacional de Pesquisa em Energia e Materiais (CNPEM)
(19) 3517.5105
sergio.carrare@cnpem.br
www.cnpem.br

Dennis Massarotto Campos
Gerente - Tecnologia de Informação e Comunicação - TIC
Centro Nacional de Pesquisa em Energia e Materiais - CNPEM
Fone +55 19 3512-1078 / Celular +55 19 99601-8625
E-mail / Skype dennis.campos@cnpem.br
www.cnpem.br

7. CONTATO PARA QUESTÕES COMERCIAIS

Gabriela Ribeiro Radomile
Suprimentos – Projeto Sirius
Centro Nacional de Pesquisa em Energia e Materiais (CNPEM)
(19) 3512.3537
gabriela.radomile@lnls.br
www.cnpem.br