18º WRNP

Workshop RNP

15 | 16 MAIO

Belém | PA

Evoluções na arquitetura para suporte a 100Gbps e outras atualizações

Fernando Frota Redigolo

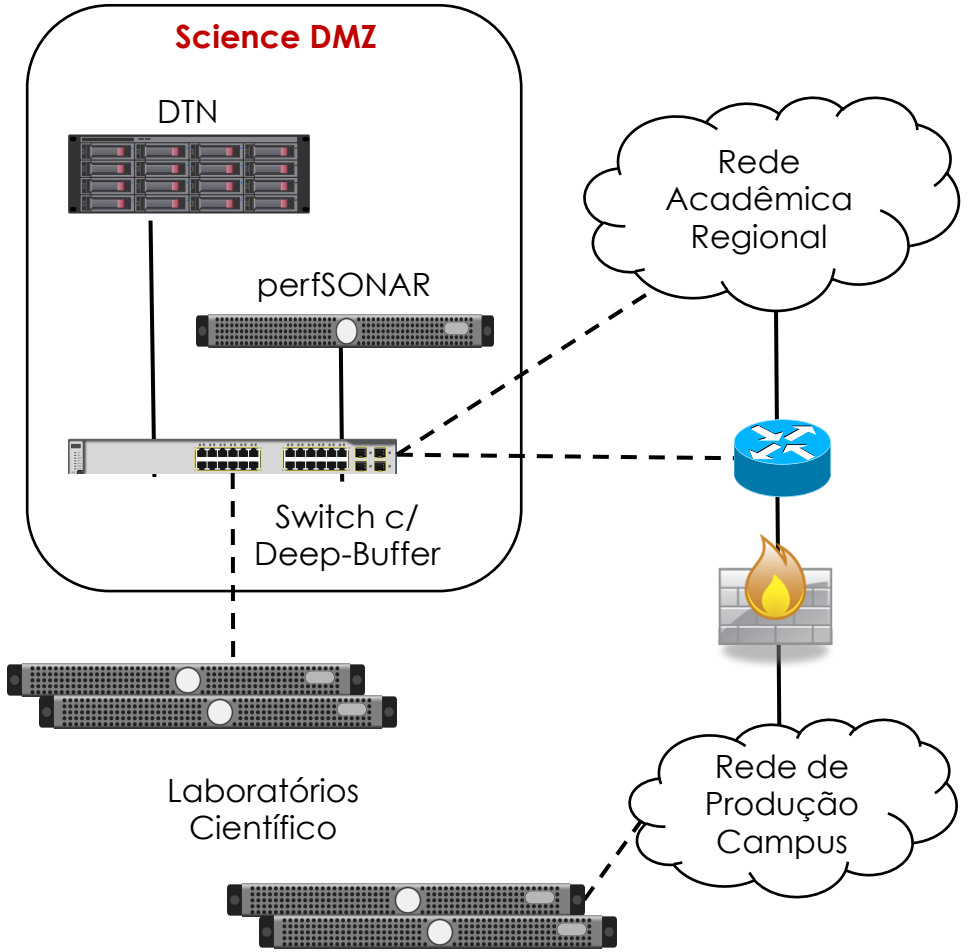
Universidade de São Paulo

LARC-USP

Evoluções na arquitetura para suporte a 100Gbps e outras atualizações

Projeto Science DMZ (2013 - Atual)

- Investigação e Apoio para implantação de infraestrutura de rede especializada para aplicações científicas
  - DMZs Científicas
- Conectividade / Segurança / Monitoramento diferenciados
- Principal Aplicação: Transferência de arquivos grandes, na ordem de GBs e **TBs**.
- Implementada em **9 Instituições:**
  - USP, RNP, CPTEC/INPE
  - CBPF, UFRJ, IFPE, UFPE
  - LNLS, LNCC (**PADEX**)



Cenários identificados para evolução do Projeto

Cenário 1: Demandas Científicas >>> Banda da Instituição

- Transferência Disco-a-Disco para 100 Gbps

Cenário 2: Interligação em Rede de Equipamentos Científicos Especializado

- Mini-Science DMZ

Cenário 3: Science DMZ sob Demanda

- Science DMZ-como-Serviço (Sc.DMZ-aaS)

Cenários identificados para evolução do Projeto

Cenário 1: Demandas Científicas >>> Banda da Instituição

- Transferência Disco-a-Disco para 100 Gbps

Cenário 2: Interligação em Rede de Equipamentos Científicos Especializado

- Mini-Science DMZ

Cenário 3: Science DMZ sob Demanda

- Science DMZ-como-Serviço (Sc.DMZ-aaS)

Data Transfer Node (DTN) p/ 10 Gbps (Valores Nominais)

6/12 Gbps por disco (SATA/SAS)

PCI EXPRESS

x8 v.3: 63 Gbps

n x Discos SATA/SAS (RAID-0)

Disco Linha Enterprise (*)	Max Sustentável Por disco (Gbps)	Discos Necessários (RAID-0)
HDD 7.2K rpm	2.0	5+
HDD 10K rpm	2.4	5+
HDD 15K rpm	2.4	5+
SSD SATA	4.5 / 4.0 (R/W)	3+
SSD SAS (**)	8.8 / 6.8 (R/W)	2+

** Linha Seagate – 05/2017*
*** Single Port SAS*

PCI EXPRESS

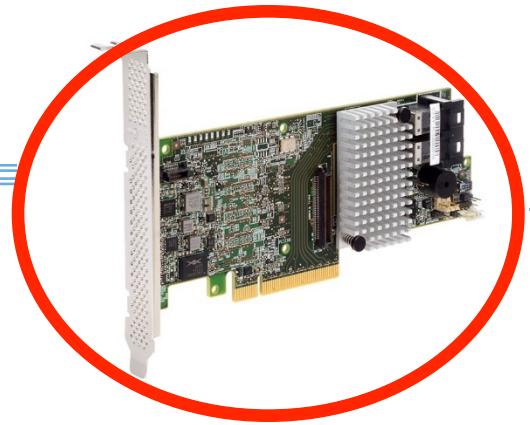
x8 v.3: 63 Gbps

DTN p/ 100 Gbps – Principais Limitações (Valores Nominais)

6/12 Gbps por disco (SATA/SAS)

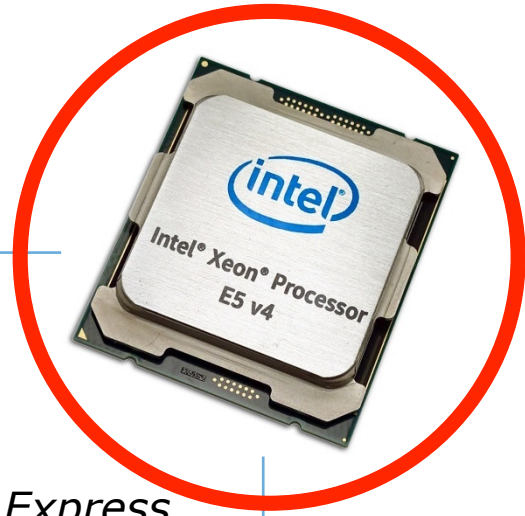
SERIAL ATA EXPRESS

Serial Attached SCSI



PCI EXPRESS

x8 v.3: 63 Gbps



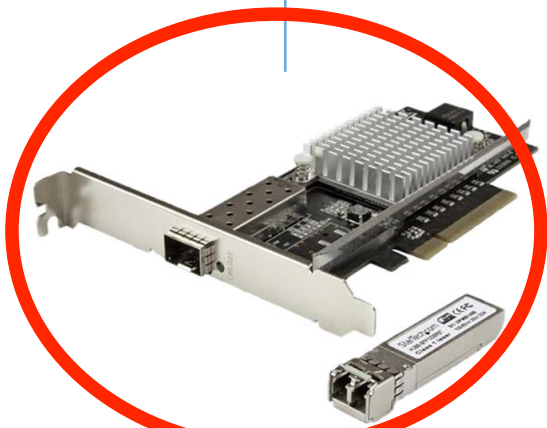
PCI Express

x8 v.3: 63 Gbps

n x Discos SATA/SAS (RAID-0)

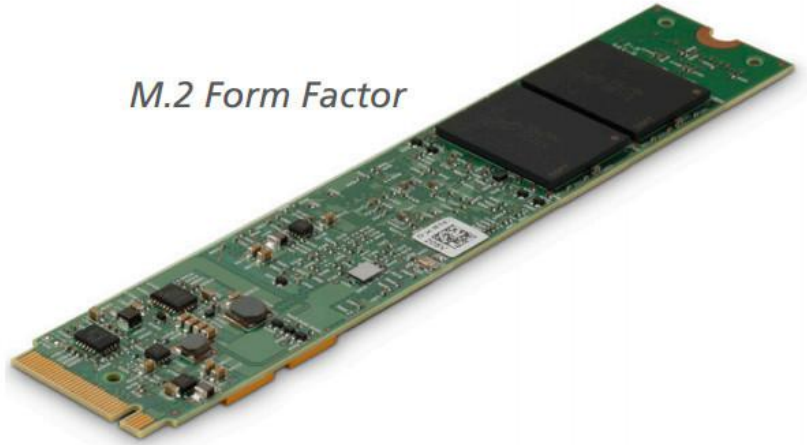
Disco Linha Enterprise (*)	Max Sustentável Por disco (Gbps)	Discos Necessários (RAID-0)
HDD 7.2K rpm	2.0	50+
HDD 10K rpm	2.4	42+
HDD 15K rpm	2.4	42+
SSD SATA	4.5 / 4.0 (R/W)	25+
SSD SAS (**)	8.8 / 6.8 (R/W)	15+

** Linha Seagate – 05/2017*
*** Single Port SAS*

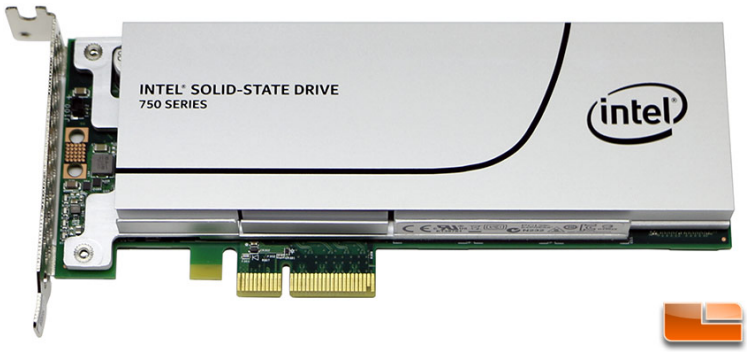


DTN 100 Gbps – Discos NVME

- SSD (*Solid State Disks*) sobre PCI Express (x4 v3)
- Menor Latência / Maior paralelismo no acesso aos dados



nvm EXPRESS



DTN 100 Gbps – Discos e Rede (Valores Nominais)

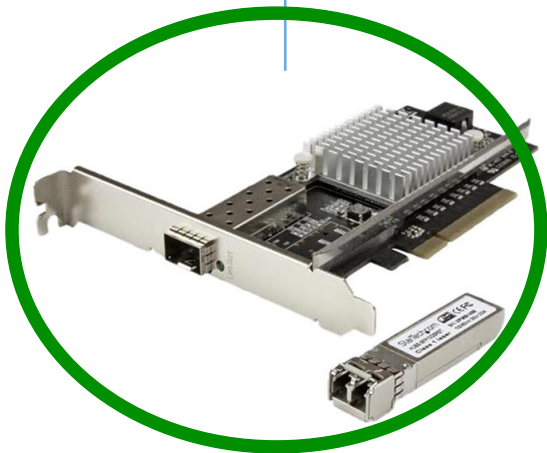
PCI Express x4 v3: ***32 Gbps / disco***

Disco Linha Enterprise	Max Sustentável Por disco (Gbps)	Discos (RAID-0)
SSD SATA (*)	4.5 / 4.0 (R/W)	25+
SSD SAS (*)	8.8 / 6.8 (R/W)	15+
SSD NVME (**)	**22.4 / 15,2 (R/W)**	**7+**

5+ Discos NVME (RAID-0)

** Linha Seagate – 05/2017, Single Port SAS*
*** Linha Intel P3700 – 05/2017*

x16 v.3: 128 Gbps
x16 v.4: 256 Gbps



- *Cuidados com Arquitetura da placa mãe – slots PCIe x16 e compartilhamento de banda*

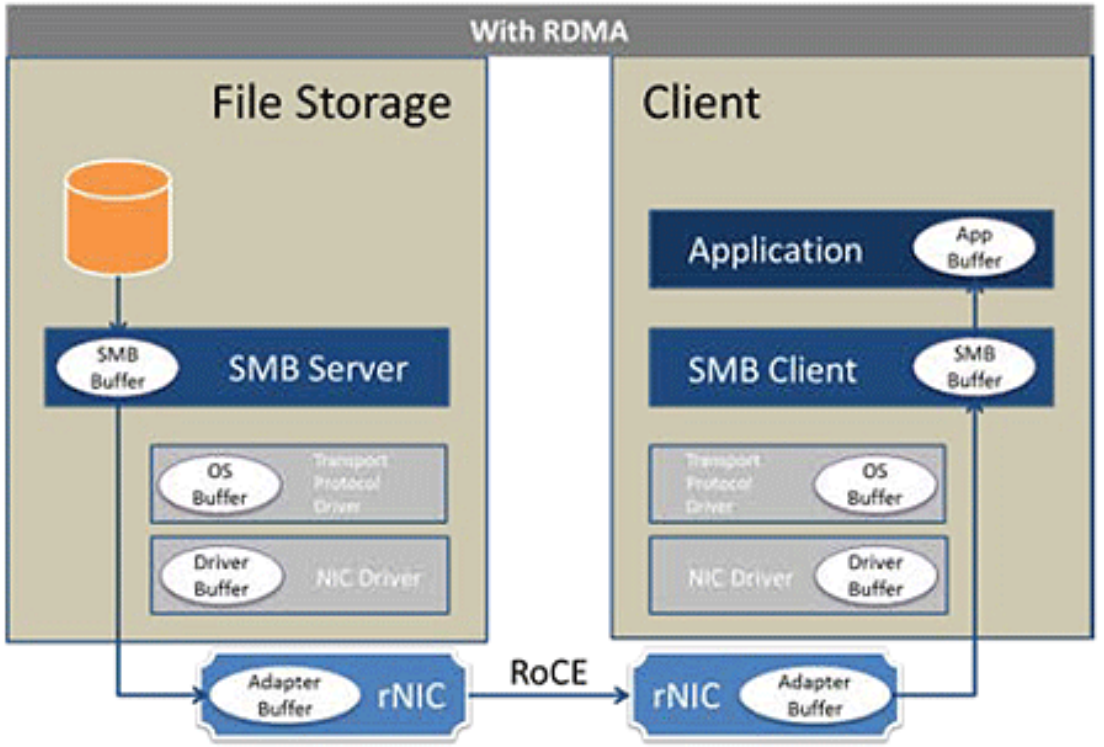
DTN 100 Gbps – CPU

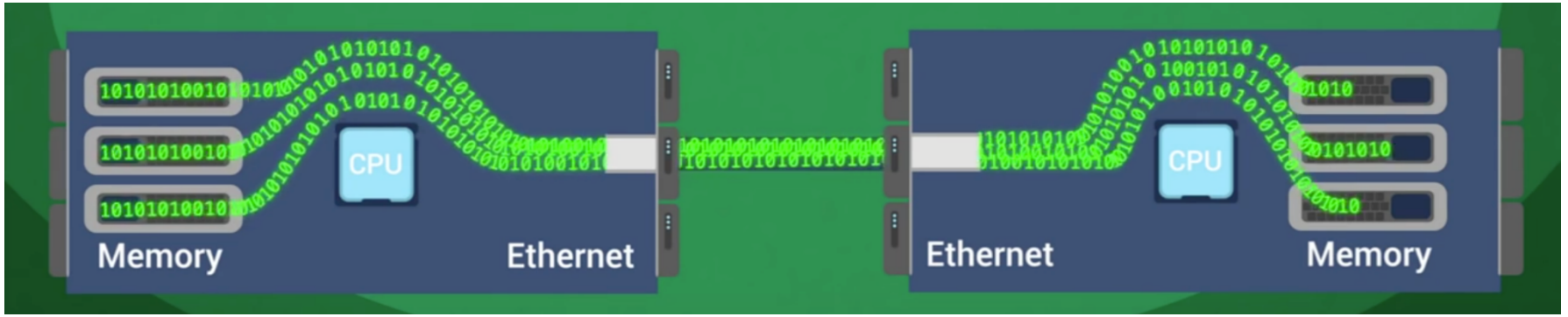
Minimizar processamento CPU

Diversas tecnologias a serem investigadas

RDMA: **R**emote **D**irect **M**emory **A**ccess

- **RoCE** – **R**emote **o**ver **C**onverged **E**thernet
- Cópia Memória-Memória s/ uso de CPU
- Necessidade de Aplicações Especializadas e Circuitos de Camada 2



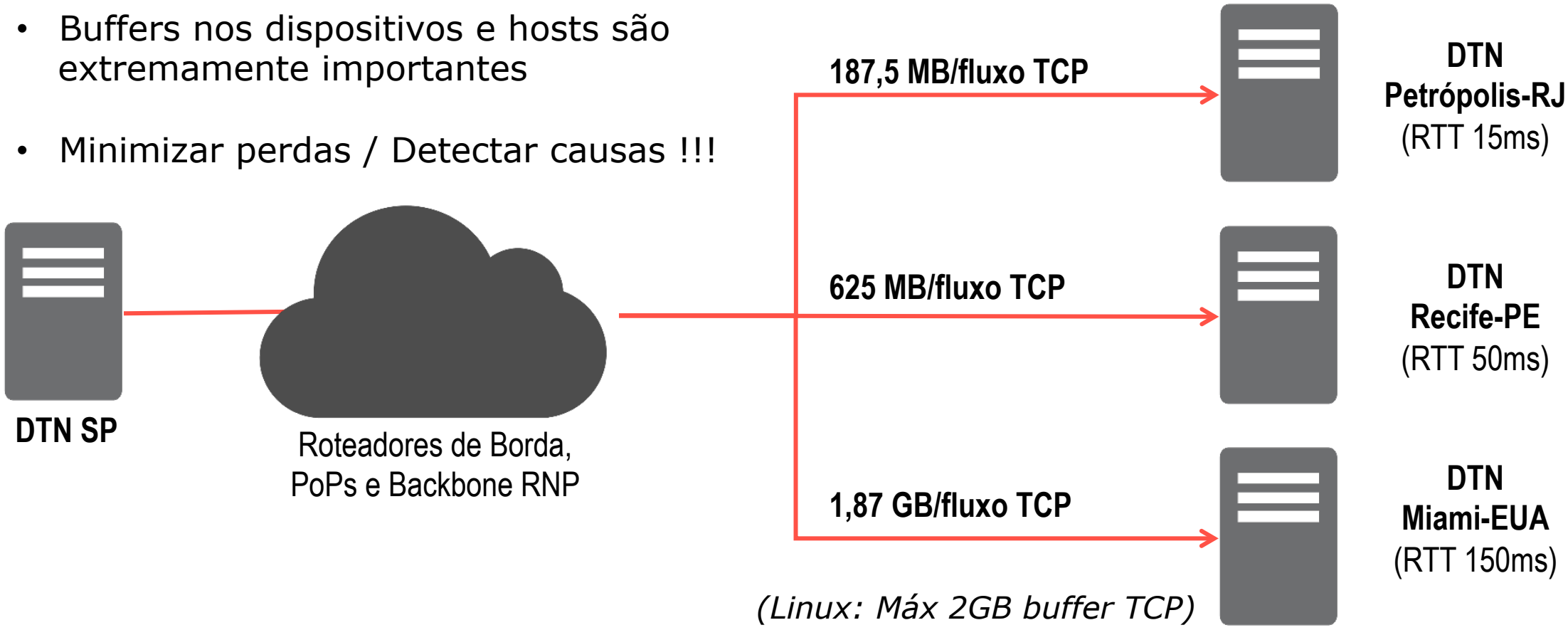


Fonte: Mellanox

100G – Rede

Perdas na rede são ainda mais **significativas** para transferências TCP a 100G – tamanho das **Janelas TCP**

- Buffers nos dispositivos e hosts são extremamente importantes
- Minimizar perdas / Detectar causas !!!



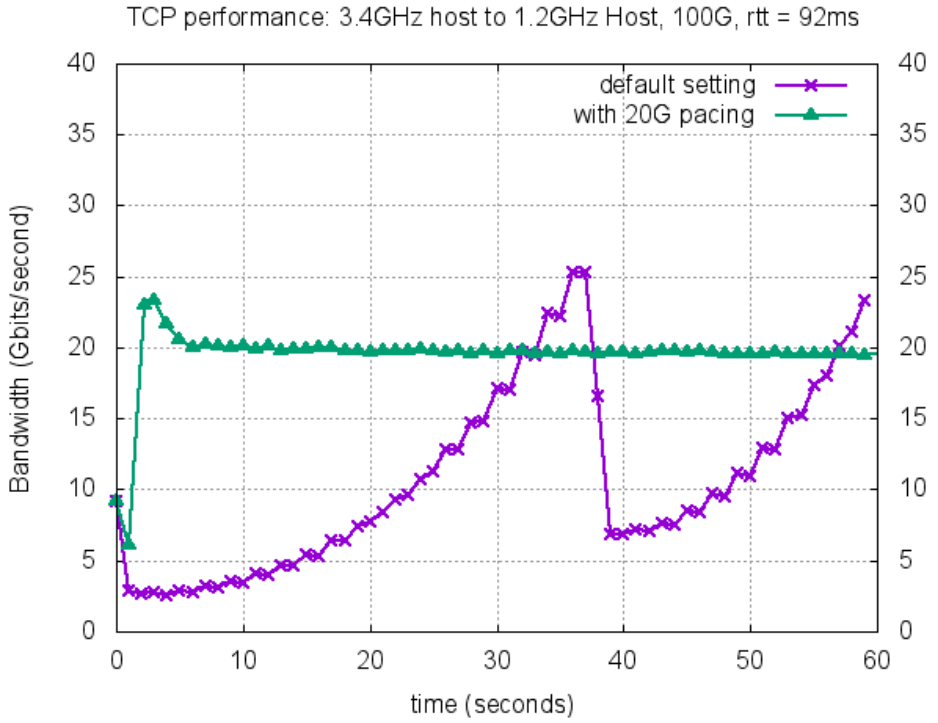
100G – Rede

Uso do *Scheduler* de *Fair Queing* (FQ) nos DTNs recomendado pela Esnet

Estabilidade e melhor desempenho quando há:

- Diferença de velocidade nas interfaces (ex.: 100G p/ 10G)
- Diferença de velocidade dos hosts
- Problemas de capacidade de equipamentos de rede no caminho (buffer, processamento)

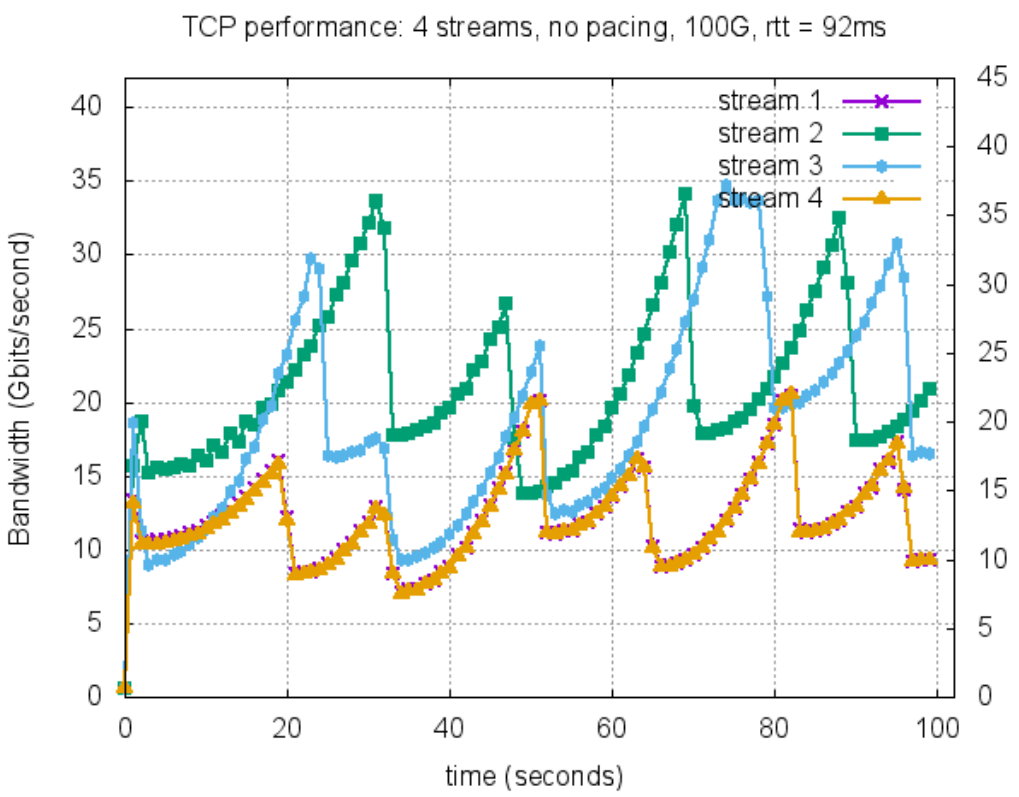
Necessário determinar manualmente a taxa transmissão para o FQ

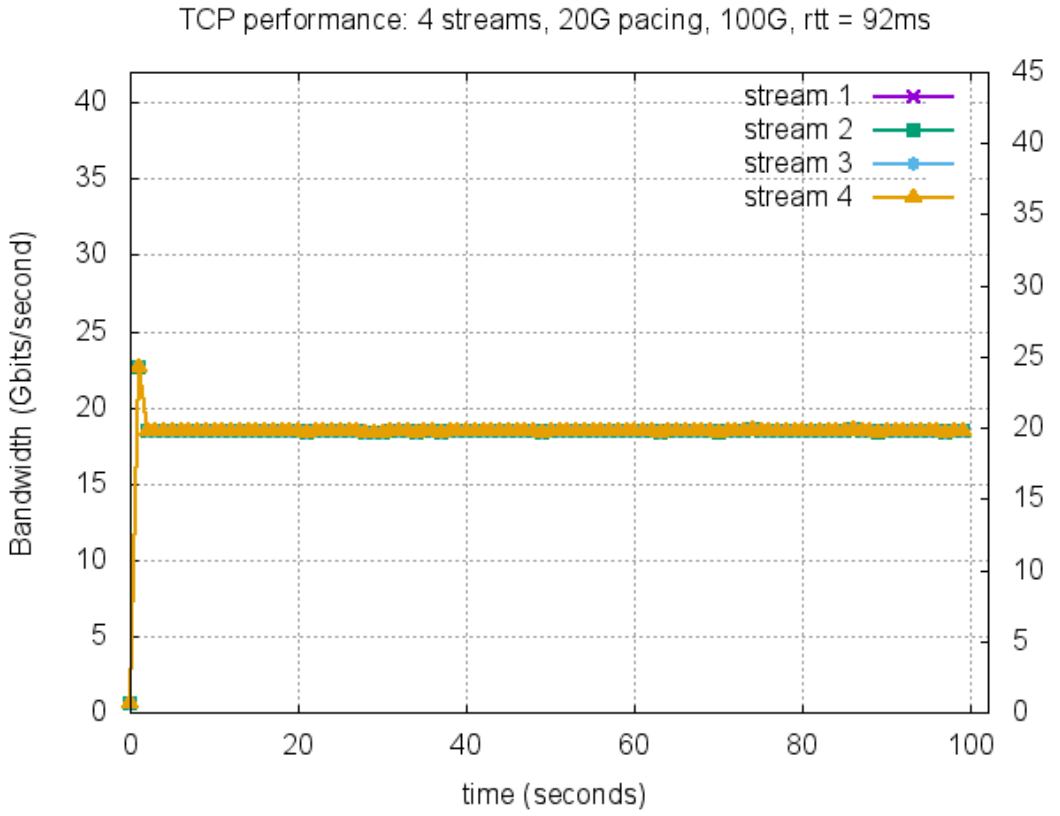


Fonte: HANFORD, N., TIERNEY, B. ***Recent Linux TCP Updates, and how to tune your 100G host.*** Apresentação Internet2 Technology Exchange, Set. 2016

100G – Rede

Uso do *Scheduler* de *Fair Queing* (FQ) nos DTNs recomendado pela Esnet





Fonte: HANFORD, N., TIERNEY, B. ***Recent Linux TCP Updates, and how to tune your 100G host.*** Apresentação Internet2 Technology Exchange, Set. 2016

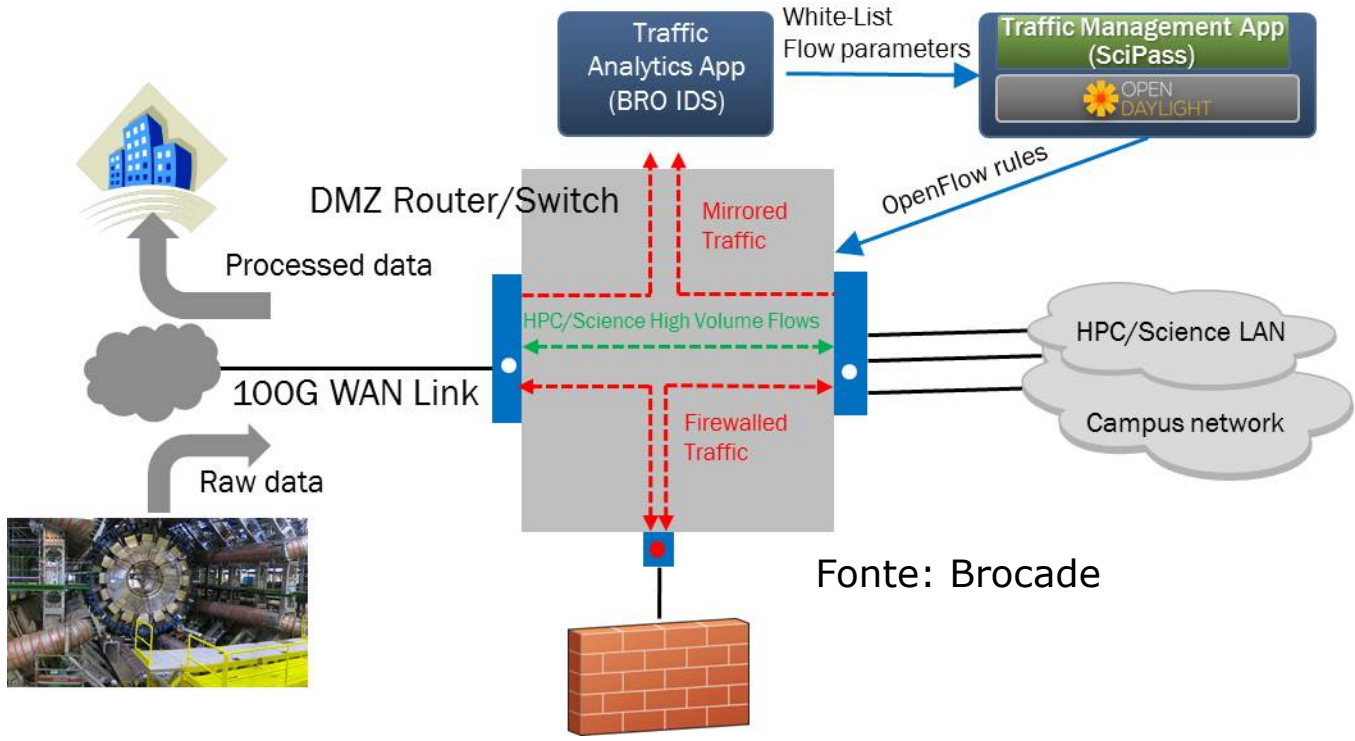
100G – Segurança

ACLs em hardware e uso de **ferramentas passivas**:

- Monitoramento: BRO (IDS em Cluster), sflow, entre outros para identificação de tráfegos científicos e mitigação de ataques;
- SDN para "*by-pass*" de firewalls para fluxos científicos;
- Ex.: SciPass, NFShunt

Outras abordagens em desenvolvimento

- Firewalls 100G para aplicações científicas



Fonte: Brocade

Cenários identificados para evolução do Projeto

Cenário 1: Demandas Científicas >>> Banda da Instituição

- Transferência Disco-a-Disco para 100 Gbps

Cenário 2: Interligação em Rede de Equipamentos Científicos Especializado

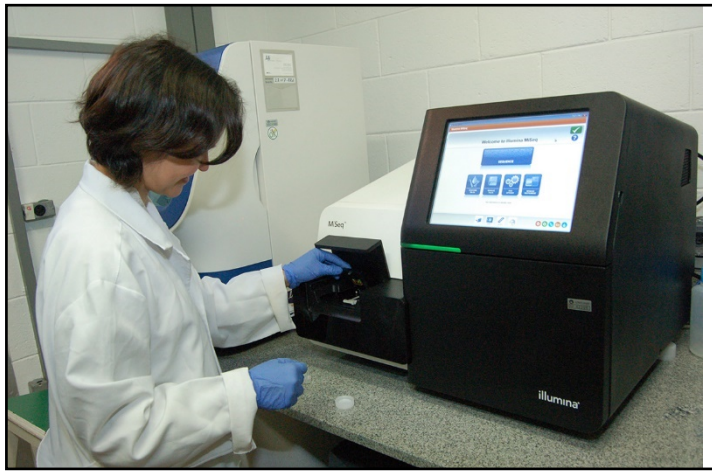
- Mini-Science DMZ

Cenário 3: Science DMZ sob Demanda

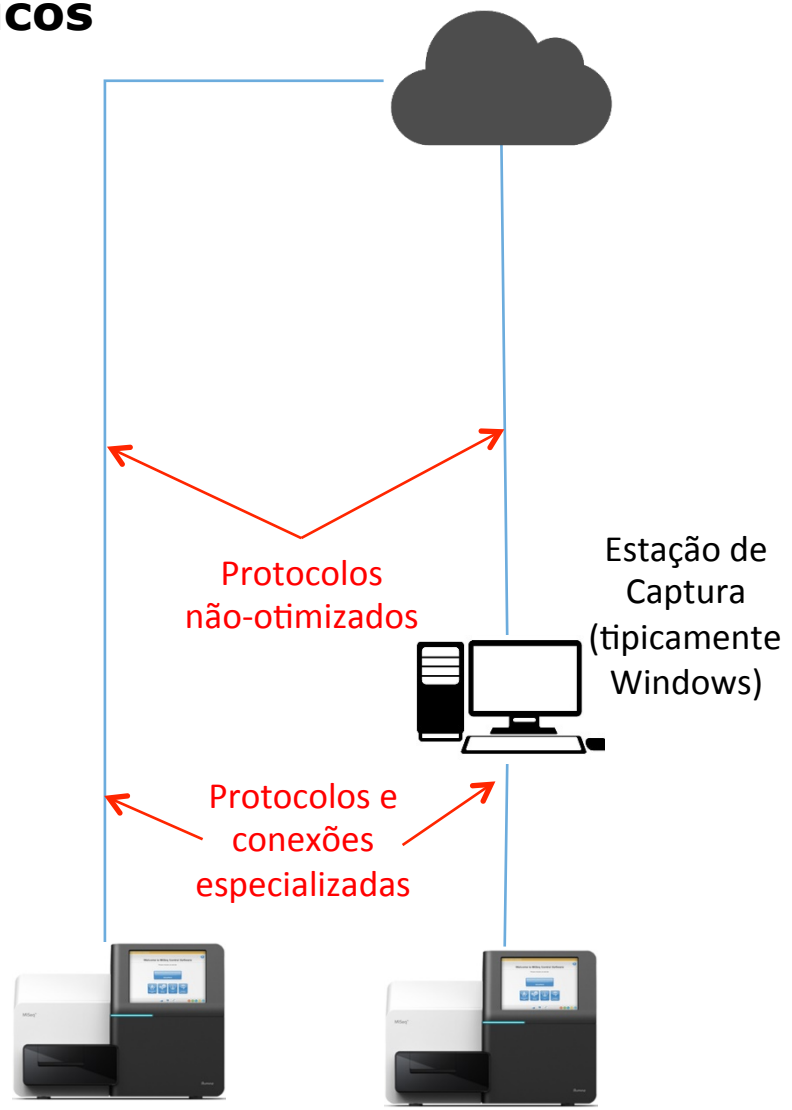
- Science DMZ-como-Serviço (Sc.DMZ-aaS)

Desafio: Interligação em Rede de Equipamentos Científicos

- Normalmente ***appliances*** especializados ou dispositivos com estação de captura Windows
- Instalados em lugares sem infraestrutura formal de TI (laboratórios, salas médicas, sites remotos, ...).



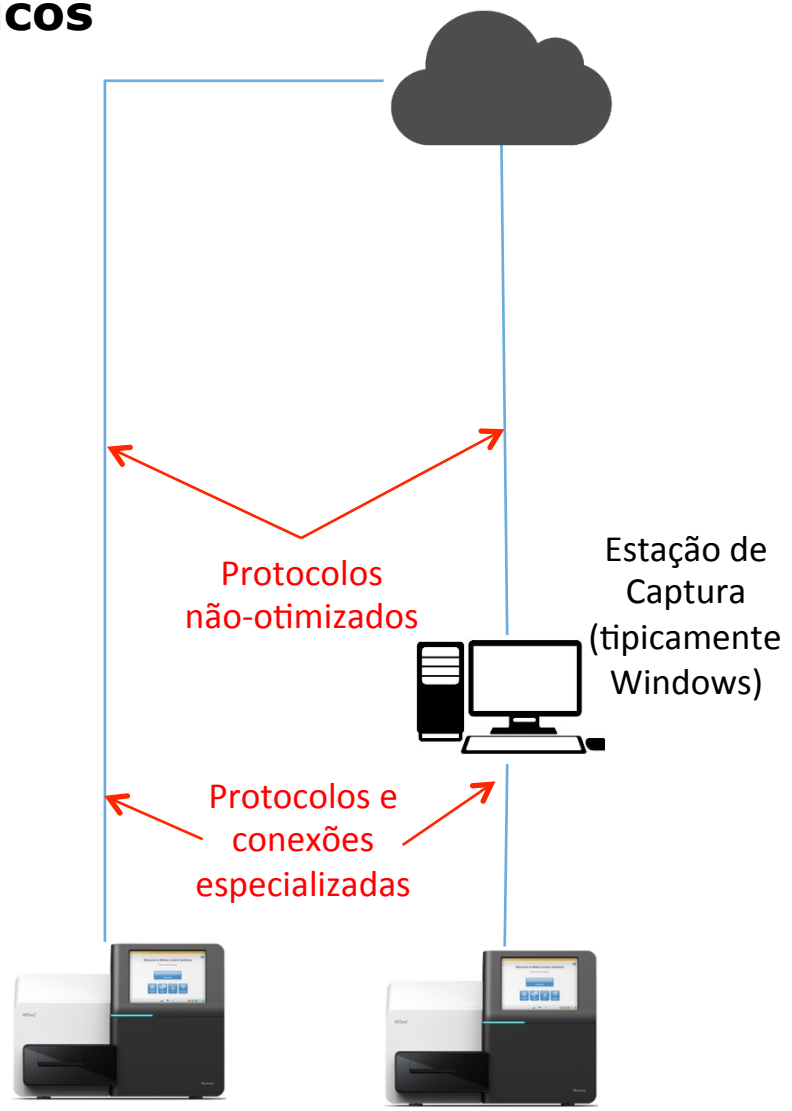
- *Sequenciador de genoma (Illumina MiSeq)*
- *Datasets de até 15GB*
- *Windows 7 interno*
- *(Foto: Unicamp)*



Evoluções na arquitetura para suporte a 100Gbps e outras atualizações

Desafio: Interligação em Rede de Equipamentos Científicos

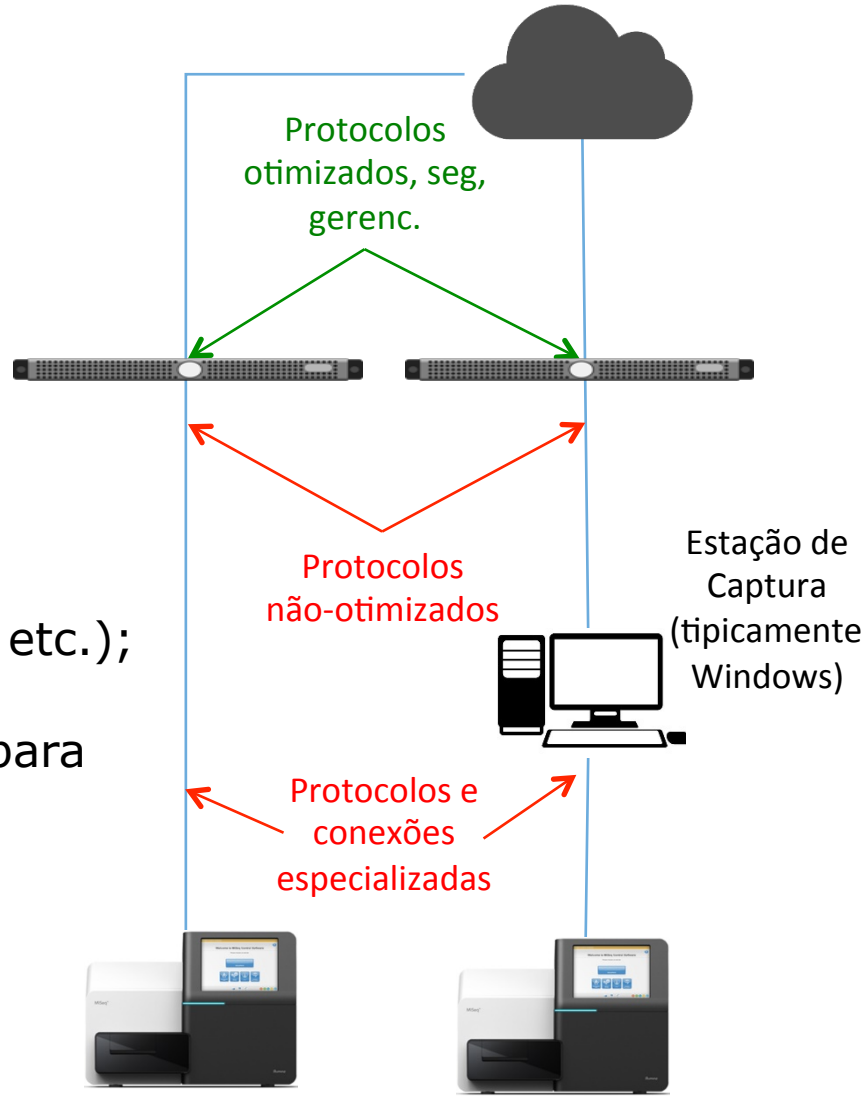
- Dificuldade de aplicar ***patches*** e instalar programas;
- Protocolos (especializados ou não) **não otimizados** para transferência de dados a longas distâncias;
- Dificuldade de **monitoramento**;
- Dificuldade de aplicar filtragem e outros mecanismos de **segurança**;



Proposta: Mini Science-DMZ

Servidor de pequeno porte englobando funcionalidades do DTN e perfSONAR, agregando Conectividade / Segurança / Monitoramento diferenciados

- Ferramentas de transferência otimizada (ex.: Globus/GridFTP);
- Ferramentas de Troubleshooting / testes do perfSONAR;
- Ferramentas de segurança *host-based* (*iptables*, *Host-based IDS*, etc.);
- Proxies de aplicação para protocolos especializados (ex.: DICOM para aplicações médicas);
- Configuração com gerenciamento centralizado;
- Proposta em Desenvolvimento na *Indiana University* (EUA);



Cenários identificados para evolução do Projeto

Cenário 1: Demandas Científicas >>> Banda da Instituição

- Transferência Disco-a-Disco para 100 Gbps

Cenário 2: Interligação em Rede de Equipamentos Científicos Especializado

- Mini-Science DMZ

Cenário 3: Science DMZ sob Demanda

- Science DMZ-como-Serviço (Sc.DMZ-aaS)

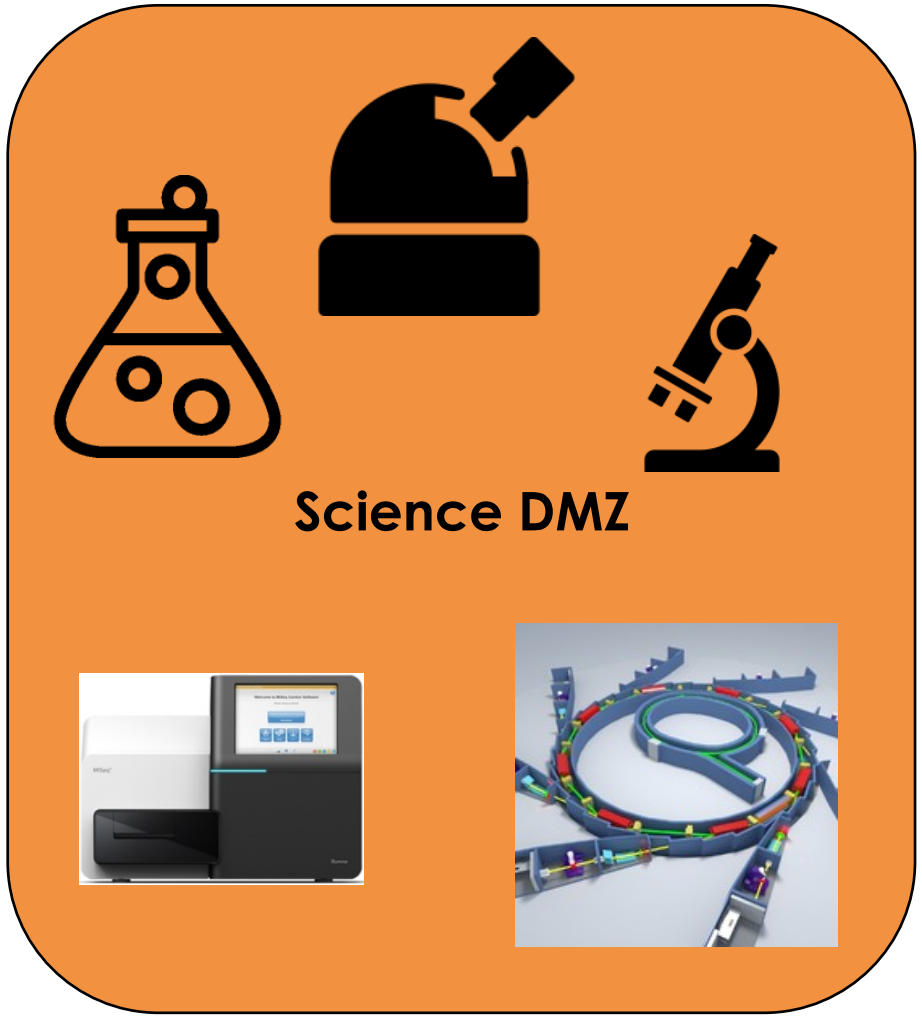
Cenário 3 – Science DMZ sob Demanda

Abordagem atual:

- **Recursos físicos dedicados** p/ máximo de desempenho
- **Otimizações Manual**

Problema: Uso por múltiplos Labs - **demandas diferenciadas**

- Requisitos mínimos de Desempenho (Otimizações Redes / SO / HW)
- Conectividade (Layer 2 / 3, ≠ latências)
- Políticas de Uso / Segurança
- Dificuldade de Gerenciamento!

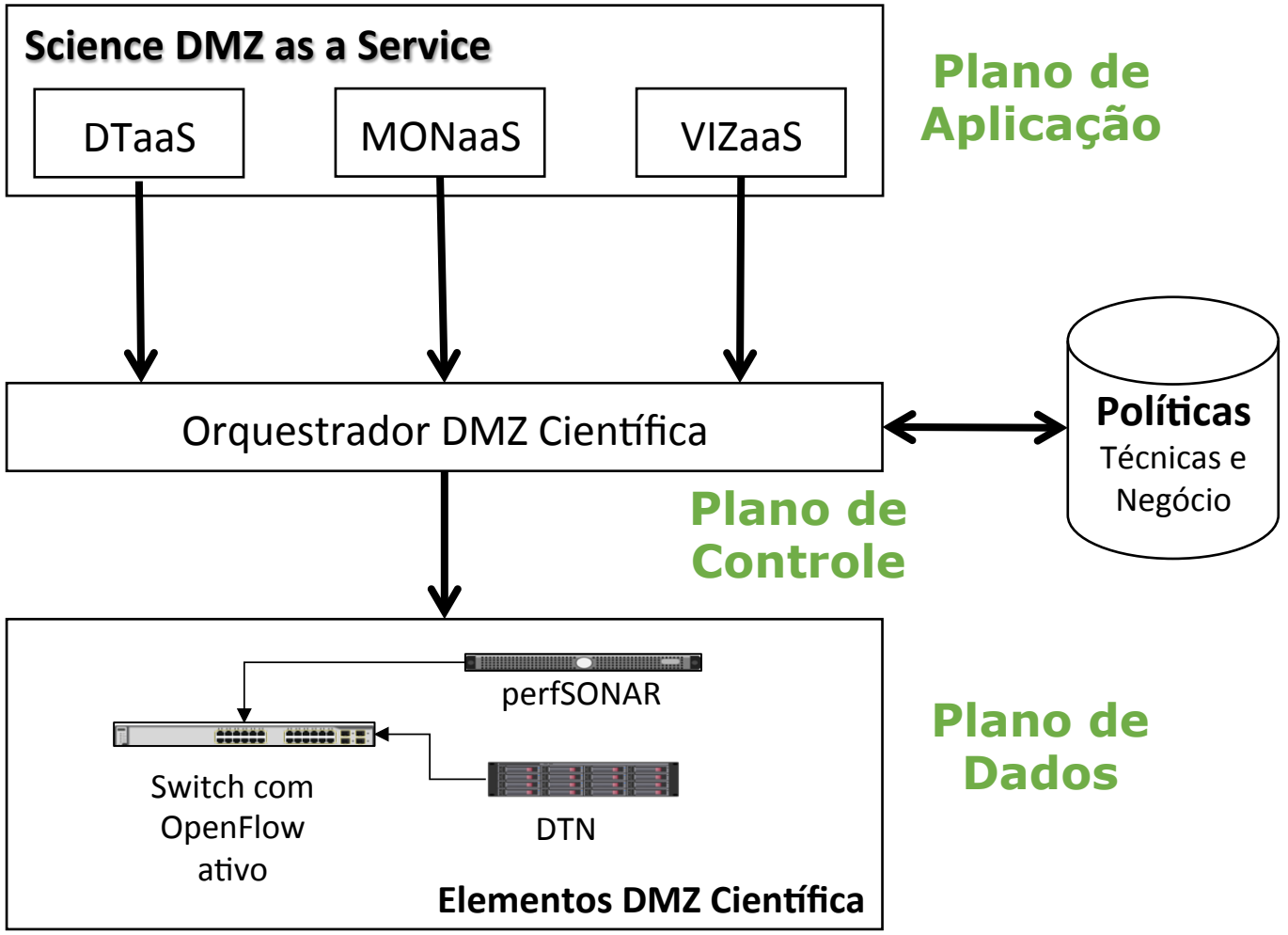


Proposta – Science DMZ-como-Serviço (Sc.DMZ-aaS)

Uso de Paradigma SDI

Software-Defined Infrastructure

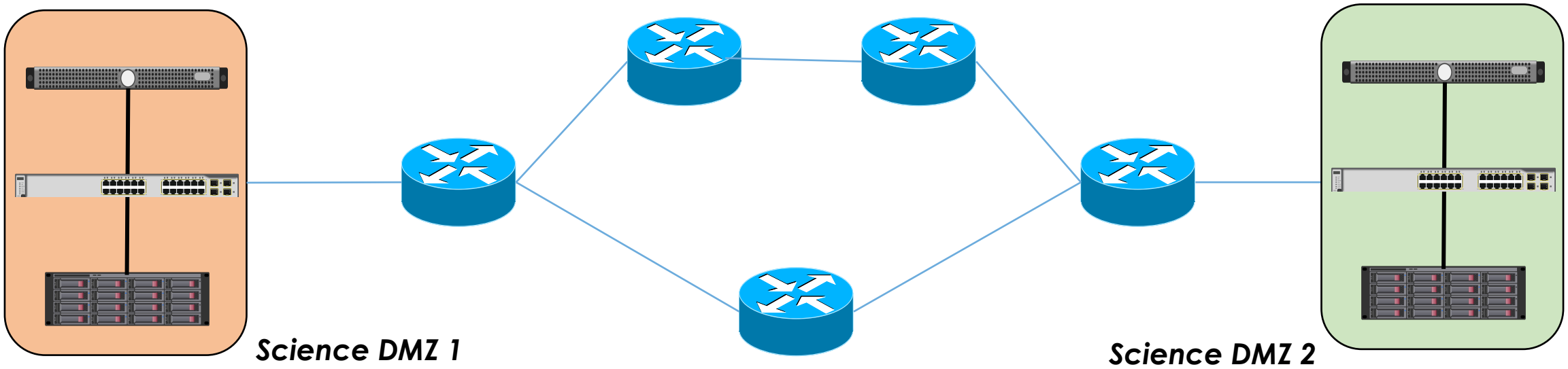
- Orquestração p/ Alocação de Recursos Físicos conforme políticas (Desempenho / Segurança / 'Negócios')
- Funções do Sc.DMZ como serviço (**NFV**)
  - *Data Transfer-as-a-Service*
  - *Monitoring-as-a-Service*
  - *Visualization-as-a-Service*



Proposta – Science DMZ-como-Serviço (Sc.DMZ-aaS)

Serviço pode ser:

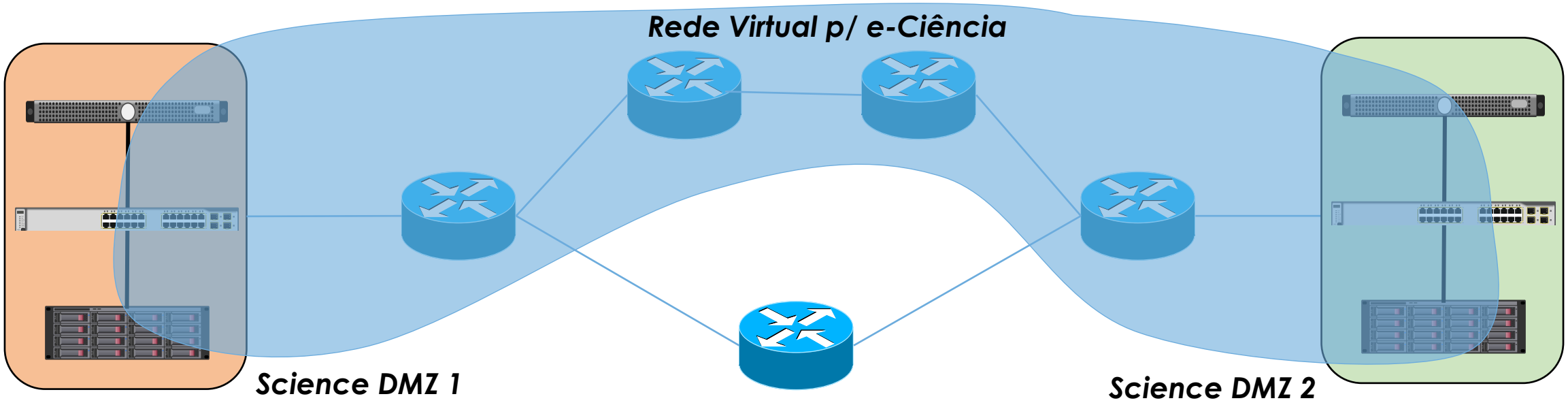
- Oferecido para os usuários de uma única instituição
- Pode ser estendido para interligação entre múltiplas Sc.DMZs → Redes Virtuais específicas p/ ciência sobre as NRENs



Proposta – Science DMZ-como-Serviço (Sc.DMZ-aaS)

Serviço pode ser:

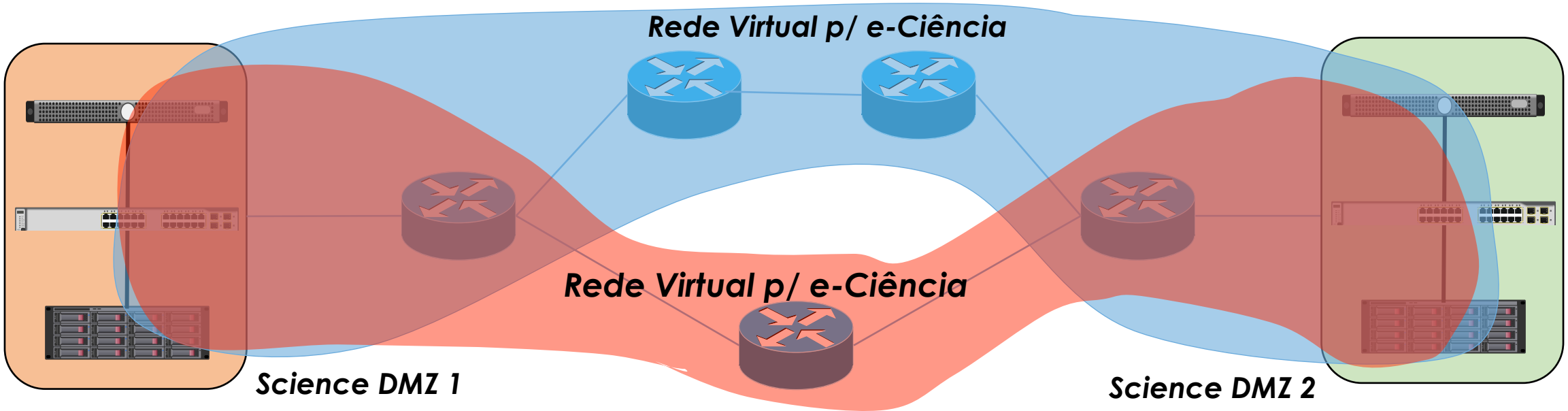
- Oferecido para os usuários de uma única instituição
- Pode ser estendido para interligação entre múltiplas Sc.DMZs → Redes Virtuais específicas p/ ciência sobre as NRENs



Proposta – Science DMZ-como-Serviço (Sc.DMZ-aaS)

Serviço pode ser:

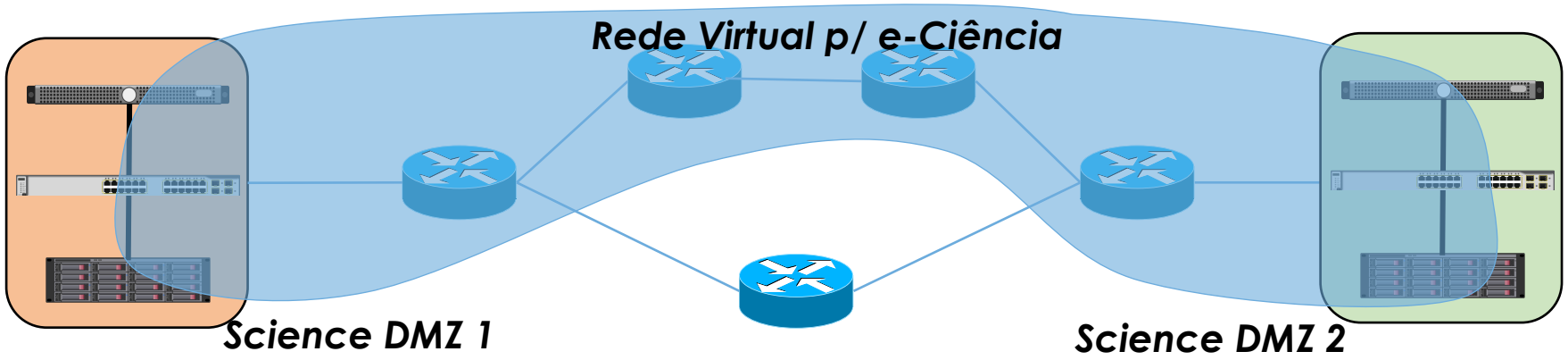
- Oferecido para os usuários de uma única instituição
- Pode ser estendido para interligação entre múltiplas Sc.DMZs → Redes Virtuais específicas p/ ciência sobre as NRENs



Proposta – Science DMZ-como-Serviço (Sc.DMZ-aaS)

Science DMZ + Projetos de Internet do Futuro (eg. FIBRE-BR, GENI-EUA)

- Provisionamento automatizado de recursos de computação / rede entre instituições (p/ experimentos de rede)
- 'Slices Científicos' otimizados → 'Virtual Sc.DMZs'
- Exemplos: SCInet (ESNet, NERSC, RENCI e ExoGENI), vSciZ (Kreonet-S), iNDIRA (OpenNSA, Globus)



18º WRNP
Workshop RNP
15 | 16 MAIO
Belém | PA

RNP
MINISTÉRIO DA DEFESA
MINISTÉRIO DA CULTURA
MINISTÉRIO DA SAÚDE
MINISTÉRIO DA EDUCAÇÃO
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
BRASIL GOVERNO FEDERAL

Obrigado!

Fernando Frota Redigolo

fernando@larc.usp.br