MonIPÊ: Serviço de Medições de Desempenho de Redes

Rede Nacional de Ensino e Pesquisa

Alex Soares de Moura – alex@rnp.br

GT-ER 39

28 a 30 de Maio de 2015, Rio de Janeiro - RJ

RNP

Ministério da **Cultura**

Ministério da **Saúde**

Ministério da **Educação**

Ministério da **Ciência, Tecnologia e Inovação**

Monitorar o desempenho das redes é fundamental para a eficiência das comunicações e aplicações

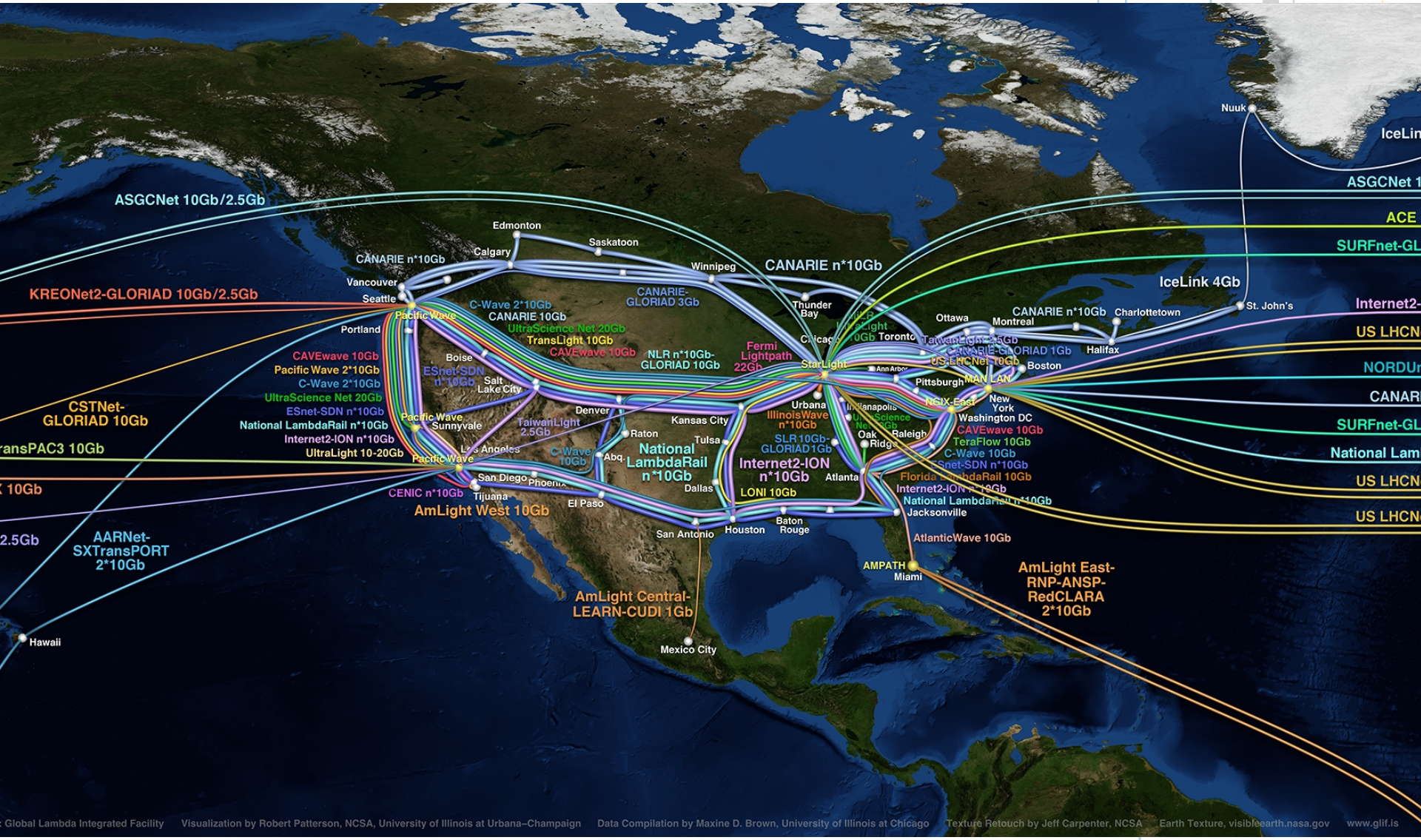
Com o Serviço MonIPÊ – compatível com o padrão perfSONAR – é possível realizar, com alta precisão, o monitoramento fim a fim do desempenho de redes

Monitoramento melhor = Redes melhores

O Problema

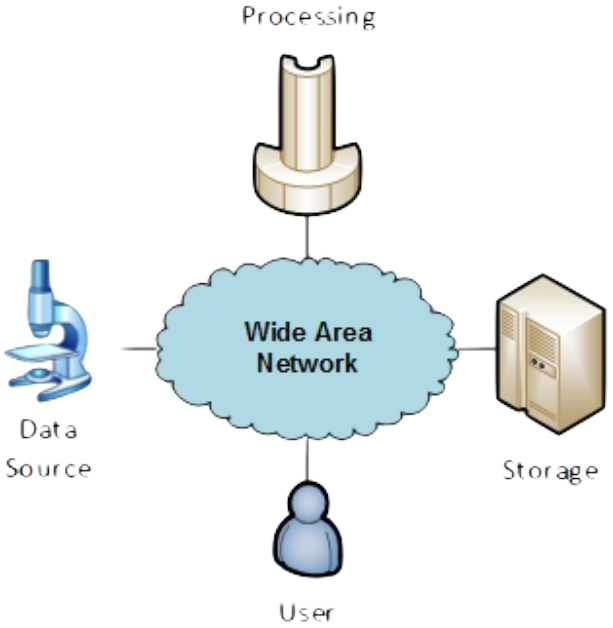
- O ecossistema de redes de Pesquisa e Educação é composto de centenas de redes internacionais, nacionais, regionais e de âmbito local
- Enquanto interconectadas, cada rede é operada por diferentes organizações (“domínios”) com diferentes políticas, clientes, modelos de financiamento, hardware, largura de banda e configurações.

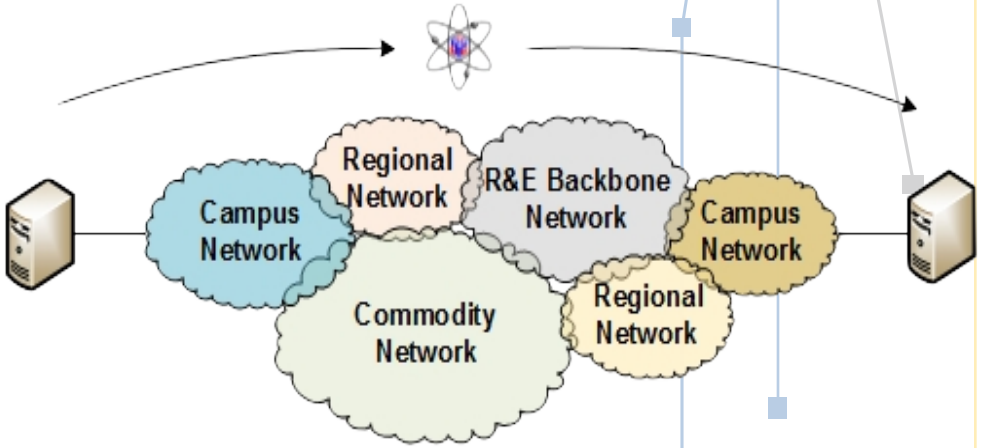
O Problema


ASGCNet 10Gb/2.5Gb
KREONet2-GLORIAD 10Gb/2.5Gb
CSTNet-GLORIAD 10Gb
ransPAC3 10Gb
X 10Gb
2.5Gb
AARNet-SXTransPORT 2*10Gb
Hawaii
Edmonton
Saskatoon
Calgary
CANARIE n*10Gb
Winnipeg
Vancouver
Seattle
Portland
Pacific Wave
C-Wave 2*10Gb
CANARIE 10Gb
UltraScience Net 20Gb
TransLight 10Gb
CAVEwave 10Gb
Pacific Wave 2*10Gb
C-Wave 2*10Gb
UltraScience Net 20Gb
ESnet-SDN n*10Gb
National LambdaRail n*10Gb
Internet2-ION n*10Gb
UltraLight 10-20Gb
Boise
ESnet-SDN n*10Gb
Salt Lake City
Denver
Sunnyvale
Pacific Wave
TaiwanLight 2.5Gb
Los Angeles
Pacific Wave
C-Wave 10Gb
Abq.
Raton
Kansas City
Tulsa
National LambdaRail n*10Gb
Dallas
San Diego
Phoenix
Tijuana
CENIC n*10Gb
AmLight West 10Gb
El Paso
San Antonio
Houston
Baton Rouge
AmLight Central-LEARN-CUDI 1Gb
Mexico City
CANARIE-GLORIAD 3Gb
NLR n*10Gb-GLORIAD 10Gb
Fermi Lightpath 22Gb
Thunder Bay
Chicago
StarLight
Urbana
IllinoisWave n*10Gb
SLR 10Gb-GLORIAD1Gb
Internet2-ION n*10Gb
LONI 10Gb
Atlanta
Ottawa
Toronto
Montreal
CANARIE n*10Gb
Charlottetown
St. John's
Halifax
Boston
IceLink 4Gb
US LHCNet 10Gb
Ann Arbor
Pittsburgh
MAN LAN
NGIX-East
New York
Washington DC
CAVEwave 10Gb
Oak Ridge
Raleigh
TeraFlow 10Gb
C-Wave 10Gb
ESnet-SDN n*10Gb
Florida LambdaRail 10Gb
Internet2-ION n*10Gb
National LambdaRail n*10Gb
Jacksonville
AtlanticWave 10Gb
AMPATH
Miami
AmLight East-RNP-ANSP-RedCLARA 2*10Gb
Nuuk
IceLin
ASGCNet 1
ACE
SURFnet-GL
Internet2-
US LHCN
NORDUr
CANAR
SURFnet-GL
National Lam
US LHCN
US LHCN
Global Lambda Integrated Facility
Visualization by Robert Patterson, NCSA, University of Illinois at Urbana–Champaign
Data Compilation by Maxine D. Brown, University of Illinois at Chicago
Texture Retouch by Jeff Carpenter, NCSA
Earth Texture, visibleearth.nasa.gov
www.glif.is

O Problema

- Este complexo conjunto de redes heterogêneas precisa ser operada de forma integrada, “fim a fim” para suportar colaborações de pesquisa científica distribuídas globalmente
- Na prática, problemas de desempenho são distribuídos
- Quando uma rede tem baixo desempenho, é difícil identificar a origem, e testes de rede local nem sempre são suficientes





Desempenho das redes: expectativas

- Expectativas sobre as transferências de dados
  - Em condições ideais, qual desempenho você espera da sua rede?
  - Quanto tempo demora transferir 1TB em diferentes velocidades[1] ?
- e-Ciência requer **pesquisa colaborativa**, transferências de **volumes de dados crescentes** e tem como requisito **baixo tempo de transferência**[2]
- Pesquisadores realizam trabalhos científicos colaborativos e **compartilham dados e recursos computacionais**[2]
- Com essa colaboração, **é necessário movimentar grandes volumes de dados**, da ordem de **gigabytes** ou até mesmo **terabytes** por dia [2]

TRANSFERÊNCIA DE 1TB

Rede (máx.)	Duração
10 Mbps	300h (12,5 dias)
100 Mbps	30h
1 Gbps	3h
10 Gbps	20min

(valores aproximados baseados na vazão máxima teórica de cada rede)

1. Fonte: ESnet Fasterdata - http://fasterdata.es.net/fasterdata-home/requirements-and-expectations/
2. Fonte: RNP Projeto Science DMZ (WRNP 2014): http://indico.rnp.br/getFile.py/access?contribId=31&resId=0&materialId=slides&confId=188

Desempenho das redes: expectativas (2)

Vazão de Dados no Tempo

- Quanto tempo demora para transferir N bytes (em bits/s)

Dados	1min	5min	20min	1h	8h	24h	7d	30d
1XB					277,78 Tbps	92,59 Tbps	13,23 Tbps	3,09 Tbps
100PB					27,78 Tbps	9,26 Tbps	1,32 Tbps	308,64 Gbps
10PB	1.333,33 Tbps	266,67 Tbps	66,67 Tbps	22,22 Tbps	2,78 Tbps	925,93 Gbps	132,28 Gbps	30,86 Gbps
1PB	133,33 Tbps	26,67 Tbps	6,67 Tbps	2,22 Tbps	277,78 Gbps	92,59 Gbps	13,23 Gbps	3,09 Gbps
100TB	13,33 Tbps	2,67 Tbps	666,67 Gbps	222,22 Gbps	27,78 Gbps	9,26 Gbps	1,32 Gbps	308,64 Mbps
10TB	1,33 Tbps	266,67 Gbps	66,67 Gbps	22,22 Gbps	2,78 Gbps	925,93 Mbps	132,28 Mbps	30,86 Mbps
1TB	133,33 Gbps	26,67 Gbps	6,67 Gbps	2,22 Gbps	277,78 Mbps	92,59 Mbps	13,23 Mbps	3,09 Mbps
100GB	13,33 Gbps	2,67 Gbps	666,67 Mbps	222,22 Mbps	27,78 Mbps	9,26 Mbps	1,32 Mbps	0,31 Mbps
10GB	1,33 Gbps	266,67 Mbps	66,67 Mbps	22,22 Mbps	2,78 Mbps	0,93 Mbps	0,13 Mbps	0,03 Mbps
1GB	133,33 Mbps	26,67 Mbps	6,67 Mbps	2,22 Mbps				
100MB	13,33 Mbps	2,67 Mbps	0,67 Mbps	0,22 Mbps				

LEGENDA
Requer vazão abaixo de 100Mbps
Requer vazão entre de 100Mbps e 10Gbps
Requer vazão entre de 10Gbps e 100Gbps
Requer vazão acima de 100Gbps
Nota: Kilo, Mega etc. estão em unidades SI (ex.: 1KB = 1000 bytes)

1. Fonte: Esnet Fasterdata - http://fasterdata.es.net/fasterdata-home/requirements-and-expectations/

Protocolo TCP: ubíquo e frágil

As redes fornecem conectividade entre *hosts*

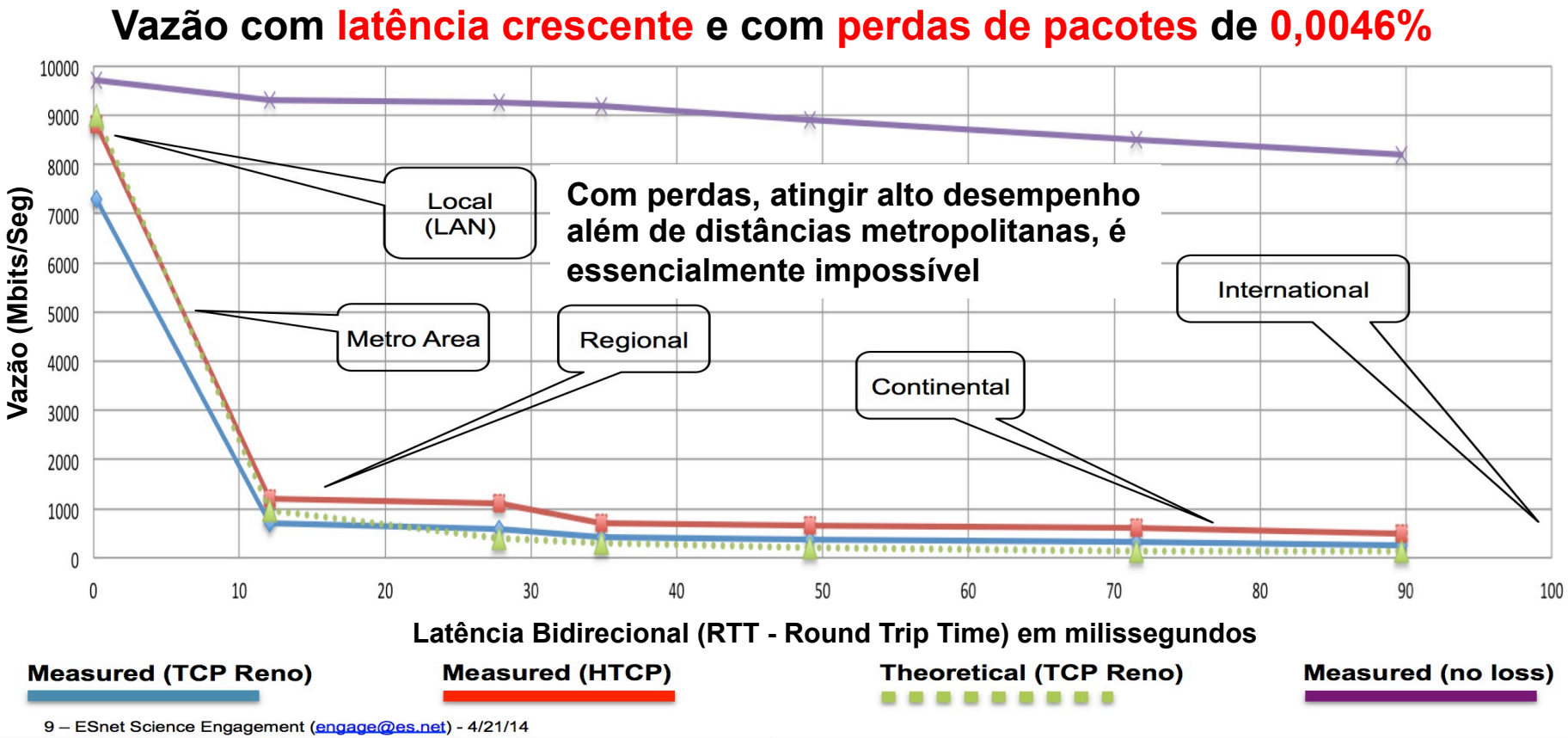
- Como aplicações e *hosts* enxergam a rede?
- Para uma aplicação, a interface para a outra extremidade é um *socket*
- A comunicação é entre aplicações, na maioria sobre TCP

- Protocolo TCP (*Transmission Control Protocol*)
  - Para o TCP, perdas de pacotes são interpretadas como congestionamentos e significam: “reduzir a taxa de transmissão”
  - Para o TCP, as perdas de pacotes, somadas à alta latência (alto RTT) causam enormes impactos no desempenho das redes
  - O TCP é o protocolo mais usado pela grande maioria das aplicações de transferências de dados (HTTP, FTP, SMTP etc)
    - Ex.: na RNP ~85% do tráfego é TCP; na Esnet, acima de 95%

Falhas completas x Falhas Parciais

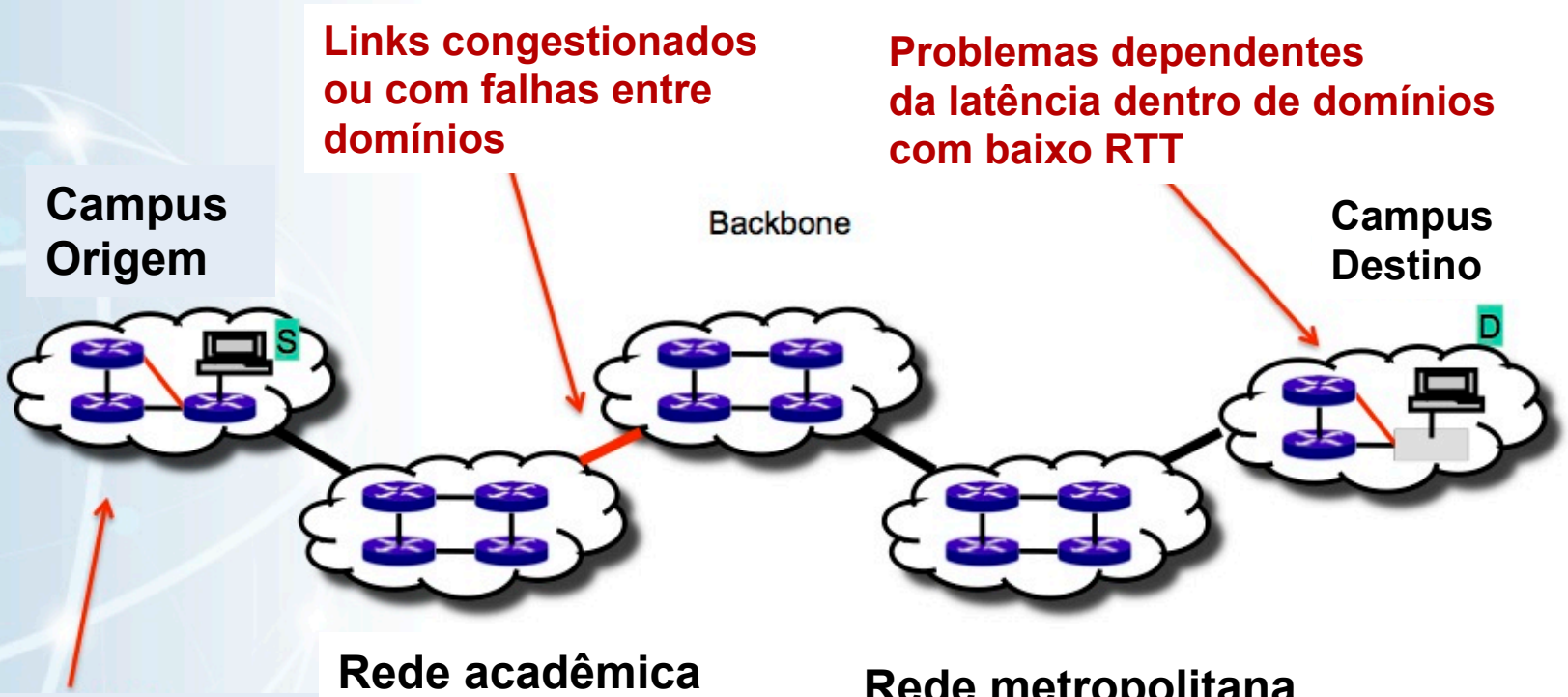
- “Falhas completas” são o tipo de problema que todos entendem
  - Cortes de fibra
  - Falha de energia
  - Hardware que para de funcionar
- Sistemas de monitoramento são bons em alertar falhas
  - Ex.: NOC percebe um alerta vermelho
  - Engenheiros são acionados
- “Falhas parciais” são diferentes e, frequentemente passam despercebidas
- Conectividade básica funciona
- Desempenho é sofrível
- Quanto devemos nos importar com falhas parciais?

Desempenho do TCP: quanto mais distante, pior. Falhas parciais causam perdas de pacotes, que o afetam

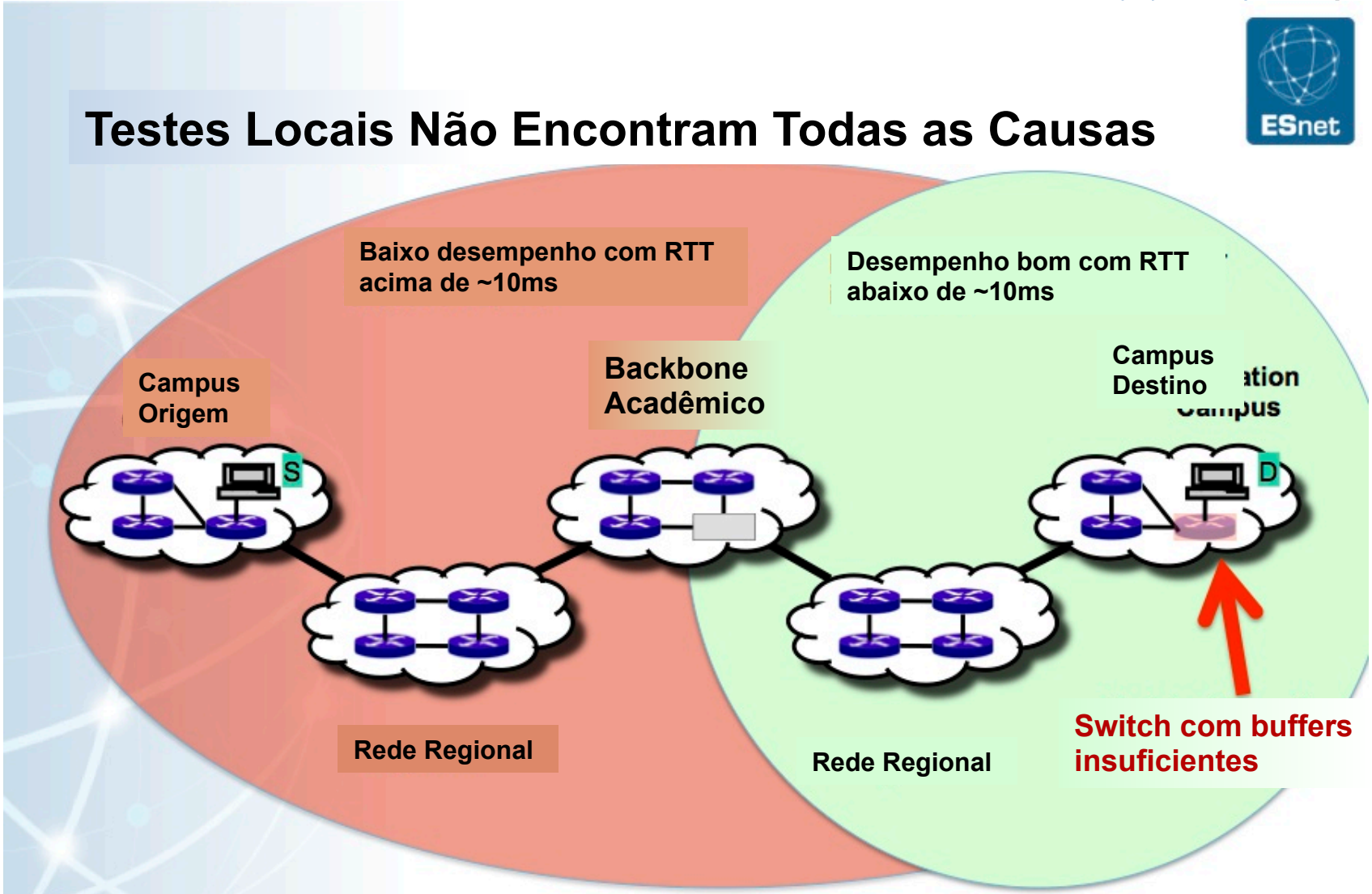


Onde estão os problemas?

Onde estão os problemas?



Testes locais não encontram todas as causas



Causas de Perdas de Pacotes

- Rede Congestionada
  - Fácil de confirmar via SNMP, “simples” de consertar com $$
  - Isto não é uma ‘falha parcial’; apenas um prob. de capacidade
  - Com frequência, pessoas assumem que congestionamento é a causa, quando na verdade não é.
- Switches com *buffers* insuficientes descartando pacotes
  - Difícil confirmar
- Firewall com processamento insuficiente descartando pacotes
  - Difícil confirmar
- Conectores ou fibras sujas, níveis ópticos falhando
  - Às vezes fácil de confirmar verificando contadores de erros
- Host sobrecarregado ou com baixo processamento perdendo pacotes
  - Fácil de confirmar olhando a carga de CPU

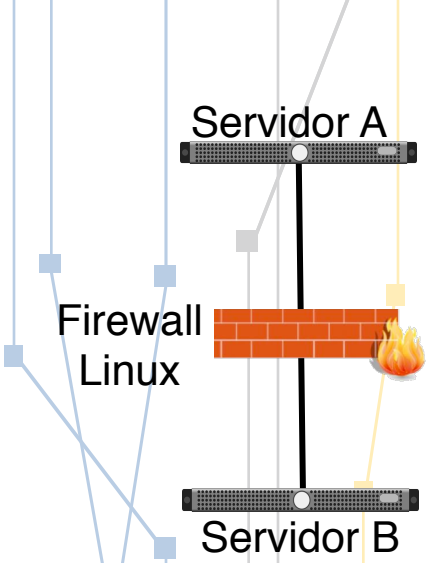
Falhas Parciais (*soft failures*)

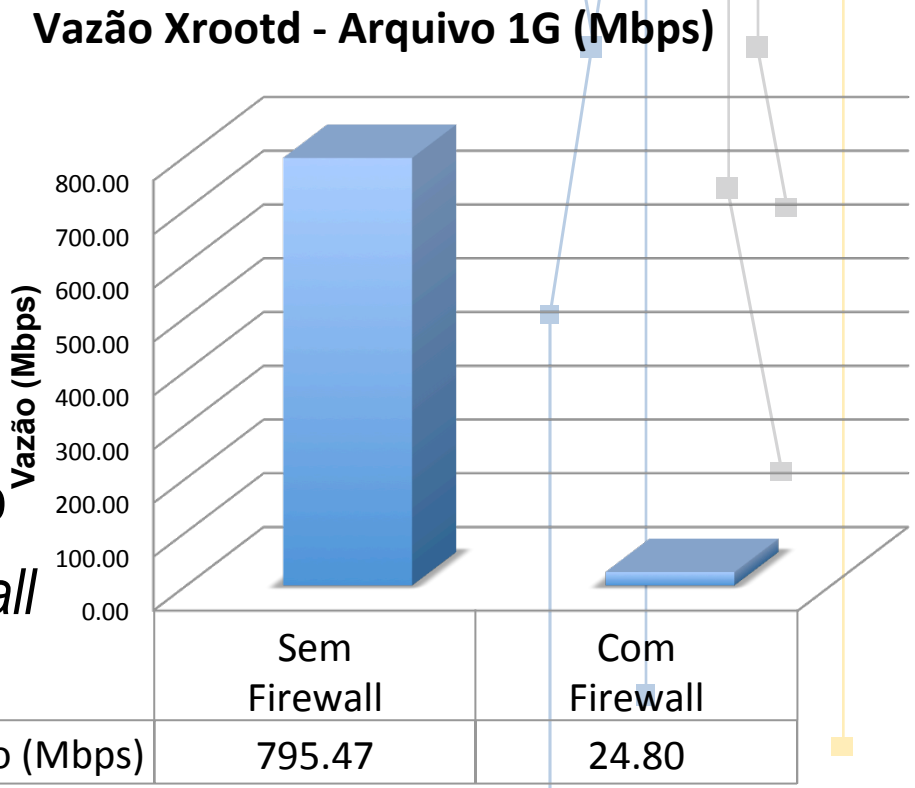
- **Falhas totais** (*hard failures*) são fáceis de detectar e consertar
- **Falhas parciais** (*soft network failures*) são falhas onde a conectividade e comunicação básica funciona, mas não é possível atingir alto desempenho.
- O TCP foi intencionalmente projetado para esconder os erros transmissões do usuário:
    - "*Enquanto o TCP funcionar adequadamente e o sistema internet não sofrer interrupções, nenhum erro de transmissão afetará os usuários.*"
      (IEN 129, RFC 716)
- Algumas falhas parciais afetam somente fluxos de alta vazão e fluxos de alta latência (RTT acima de 100ms).
- Falhas parciais podem durar por anos sem serem detectadas

Falhas Parciais: Firewalls (1)

e-Ciência e transferências de dados

- **Segurança**
- Transferência de arquivos com e sem *firewall*
- Servidores físicos dedicados sem otimização
- Transferência disco-a-disco:
  - 1 tamanho de arquivo: 1GB
  - 1 ferramenta: xrootd
- Enlaces de 1Gbps, com roteamento
- Testes com ou sem regras no *firewall*
  - 6324 regras *iptables*

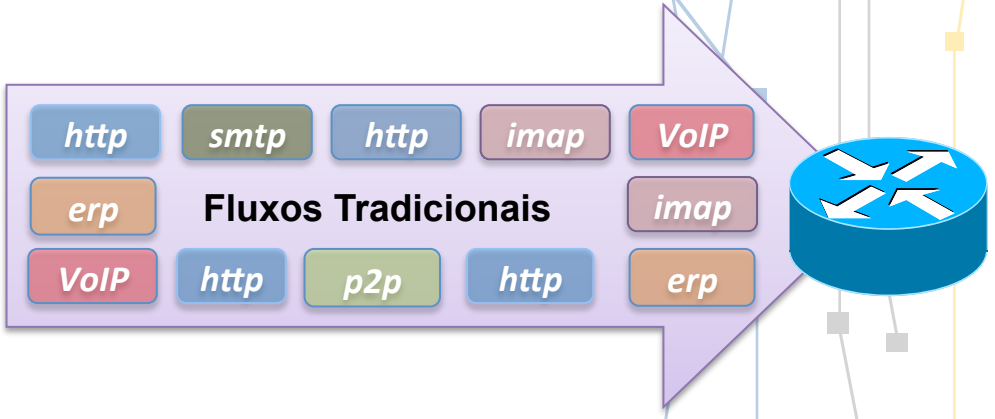




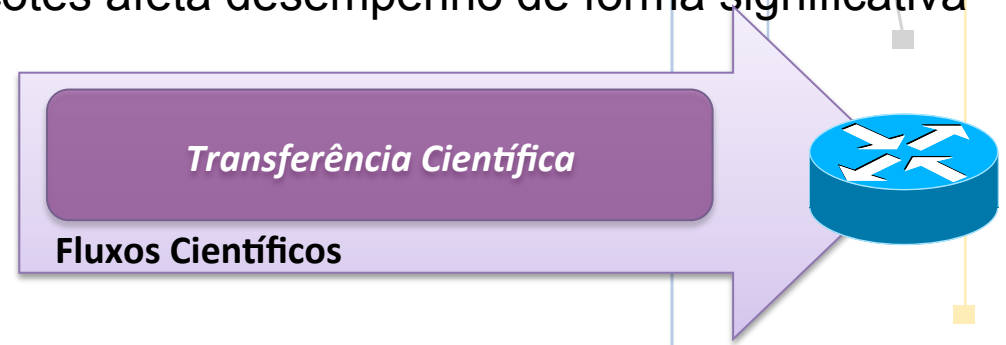
1. Fonte: RNP Projeto Science DMZ (WRNP 2014): http://indico.rnp.br/getFile.py/access?contribId=31&resId=0&materialId=slides&confId=188

Falhas Parciais: Firewalls (2)

- **Fluxos Tradicionais**
  - Grande número de fluxos consumindo pouca banda
  - Pequena taxa de perda de pacotes não afeta desempenho de forma significativa.
  - Filtragem complexa



- **Fluxos Científicos**
  - Pequeno número de fluxos consumindo muita banda
  - Pequena taxa de perda de pacotes afeta desempenho de forma significativa
  - Controle simples

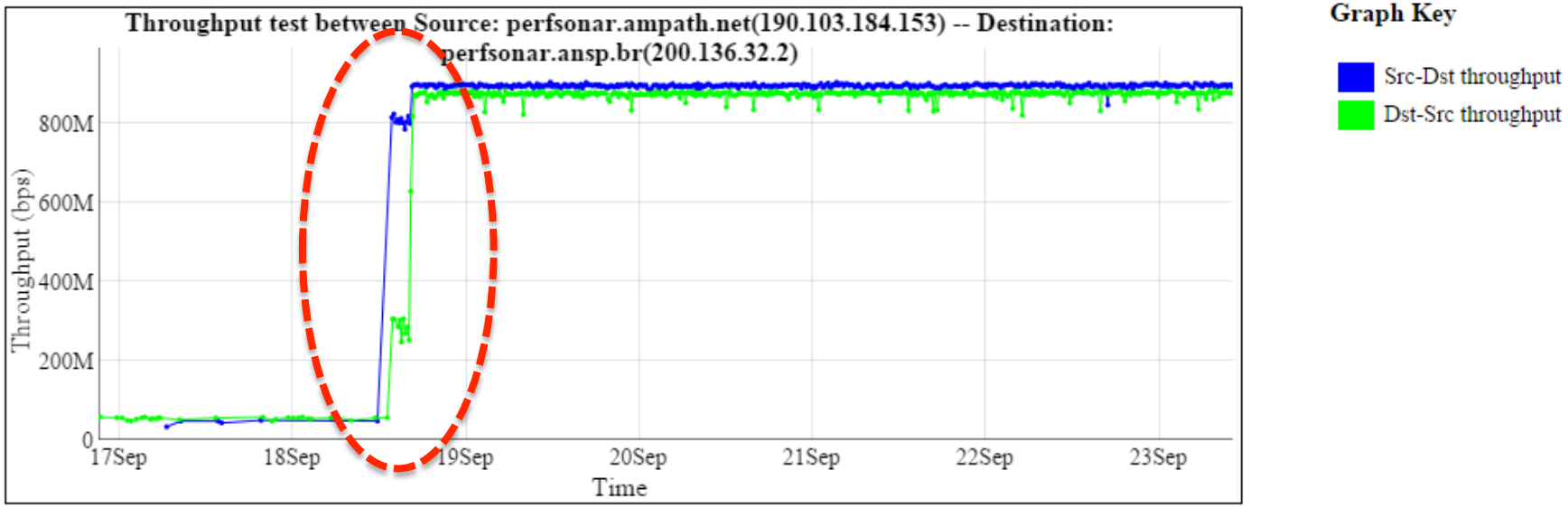


1. Fonte: RNP Projeto Science DMZ (WRNP 2014): http://indico.rnp.br/getFile.py/access?contribId=31&resId=0&materialId=slides&confId=188

Falhas Parciais: Ajustes em Hosts (1)

Desempenho

- **Vazão:** ajustes nos hosts envolvidos em comunicações e transferências de dados são necessários em caminhos com alto RTT

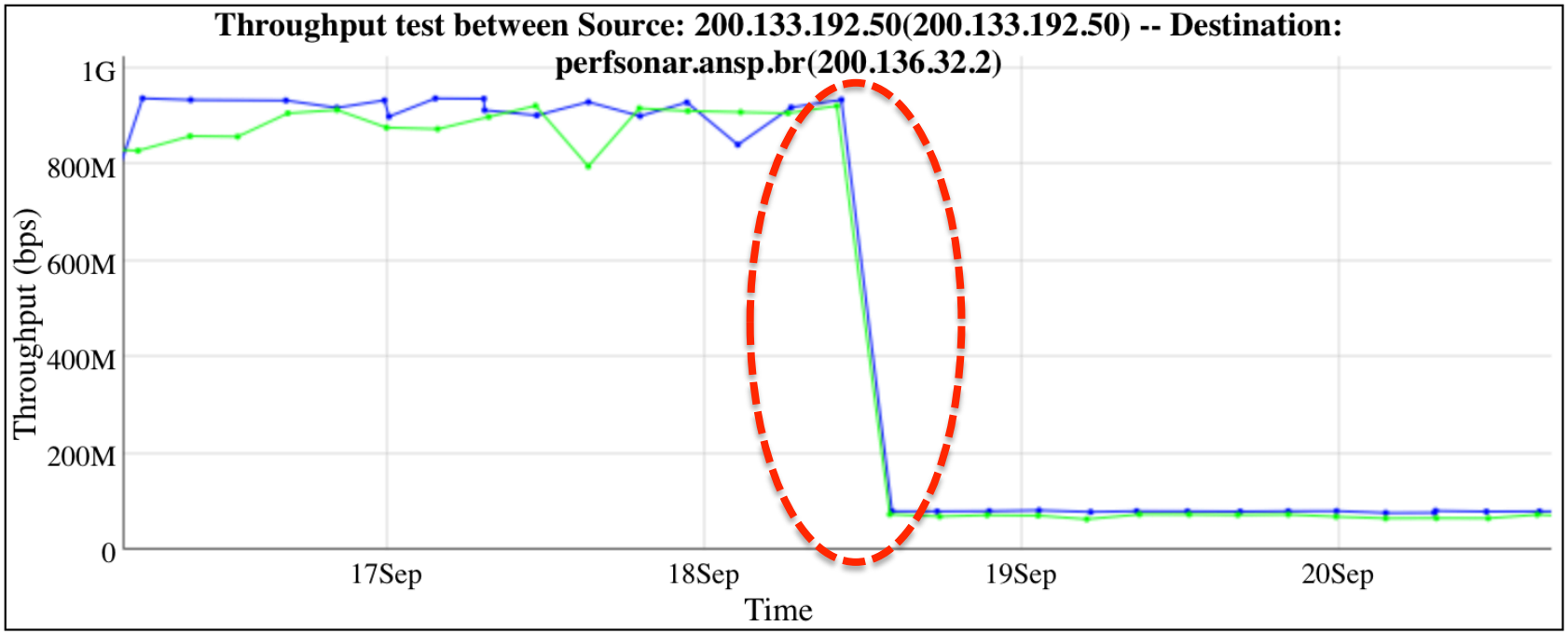


Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src > Dst	906.72M	884.68M	33.37M
Dst > Src	885.58M	814.93M	47.62M

Falhas Parciais: Ajustes em Hosts (2)

Desempenho

- Ajustes nos hosts podem ser necessários em caminhos **com baixo** RTT (~1ms)



Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src > Dst	940.22M	605.26M	75.66M
Dst > Src	938.07M	590.89M	63.7M

Métricas de Interesse

Use a ferramenta correta para o trabalho

- Para determinar a ferramenta correta, é preciso iniciar pelo que se deseja fazer...
- Que é importante medir?
  - **Perdas ou pacotes** for a de sequência, Duplicação (camada de transporte)
  - **Banda Alcançável** (p.ex.: vazão ou "*throughput*")
  - **Latência unidirecional e bidirecional** (RTT e *One Way Delay*)
  - **Variação do atraso** (*jitter, delay variation*)
  - **Utilização da interface**, **descartes**, **erros** (camada de rede)
  - **Rotas dos fluxos** (*traceroute*)
  - **MTU** (*Maximum Transmission Unit*)

Monitoramento de Rede

- Todas as redes possuem algum monitoramento.
- Atende necessidades da equipe local para entender o estado da rede
- Essa informação poderia ser útil para usuários externos?
- Essas ferramentas podem funcionar em multiplos domínios?

Além dos métodos passivos, há ferramentas para **medições ativas**.

- Frequentemente precisamos de um valor de vazão ('*throughput*').
  (É possível automatizar essa idéia?)
- Seria bom ter estatísticas de desempenho
  (Por dia? Por semana? Por ano? De múltiplos pontos?)
- Onde está o *middleware* de medições?
  (Algo que permita a fácil troca de métricas coletadas localmente, em escala global?)

perfSONAR

- Atualmente mais de 1200 pontos de medições

http://stats.es.net/ServicesDirectory/

Projeto perfSONAR

- Projeto iniciado há 10 anos nos EUA, c/ participação da GÉANT (Europa) e RNP (Brasil)
- **Motivação**
  - As redes são parte essencial das e-Ciências
  - O desempenho é fator crítico
  - Dificuldades no uso efetivo das redes WAN por cientistas
- **O perfSONAR possui:**
  - API padrão aberto
  - Camada de Web Services para comunicação entre MPs
- **O perfSONAR é uma ferramenta para:**
  - Normalizar ou compatibilizar as expectativas sobre o desempenho das redes
  - Encontrar problemas ("*soft failures*")
  - Ajudar a consertar os problemas em múltiplos domínios de rede
- Os problemas são mais difíceis quando multiplas redes são envolvidas
- perfSONAR oferece um padrão para publicação e intercâmbio de dados de monitoramentos ativos e passivos
- Esses dados são de interesse para pesquisadores e operadores de redes

Serviço MonIPÊ

Serviço de medições

de desempenho fim-a-fim entre a RNP, seus clientes, e com outras redes

Escopo e Objetivos (2015)

- Atender melhor instituições clientes da RNP;
- Estender a cobertura do monitoramento até a rede da instituição cliente;
- Colaborações em projetos de e-Ciência — p.ex: física, astronomia — e outras comunidades, colaborando com outas redes (p.ex.: ESnet, GÉANT, Internet2, NTT e RedCLARA);
- Usuários gestores de TI e comunidades de pesquisa;
- Equipamentos mais baratos;
- Simplificar a instalação;
- Melhorar a usabilidade;
- Ambiente de verificação e acompanhamento do desempenho

MEDIÇÕES

- sob demanda; temporárias e periódicas;
- de alta precisão do desempenho da rede;
- da última milha;
- **Backbone:** medições entre PoPs
- **PoPs:** medições entre PoPs e clientes diretamente conectados ao PoP
- **Internacional:** testes a outras redes acadêmicas e instituições (ex.: CERN)

MÉTRICAS

- Perdas de pacotes,
- Atraso {bi,uni}direcional
- Vazão (banda alcançável) em TCP e UDP

MonIPÊ: Componentes

Interface (GUI)

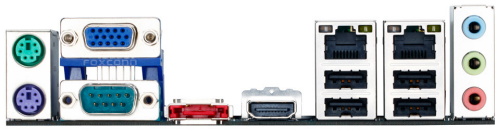
- **Portal** de Medições

Infraestrutura

- Pontos de Medição (*Measurement Points* - ***MP***)
  - **VMs e Kits Baixo Custo: até 1Gbps**
  - **Servidores dedicados: até 10Gbps**

Virtualização

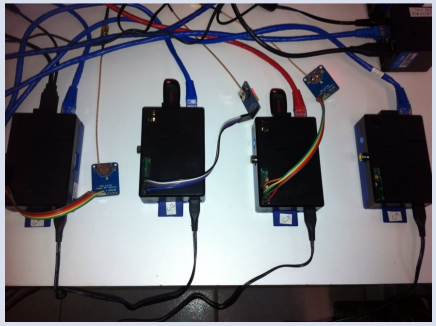
- **MPs de baixo custo**
  (kit: 1 Mini PC ou 2 SBCs[1] + GPS Adafruit)
  - 1a G. **Raspberry Pi** e **CuBox**
    CPU ARM, RAM 512MB, NIC 1GbE
  - 2a G. Mini PC **Blue Appliance 847**
    Intel Dual Core 847, 2G RAM, (2x) NICs 1GbE
- **MP 10G**
  - 1a. G. **Dell R620 - CPU**: Intel Xeon 2GHz
    RAM: 16GB, **HDD**: 2x 500GB (raid 1), NIC: 2x 10GbE + 2x 1GbE (BCM57800)



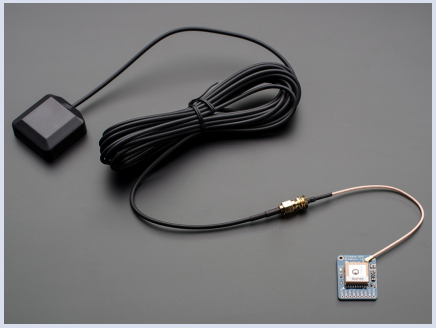
1. Single Board Computer

Kits de Baixo Custo: 1ª e 2ª geração

Hardware de 1a Geração (2013) - Custo total aproximado ~R$ 2.000,00



Raspberry Pi
Testes latência unidirecional

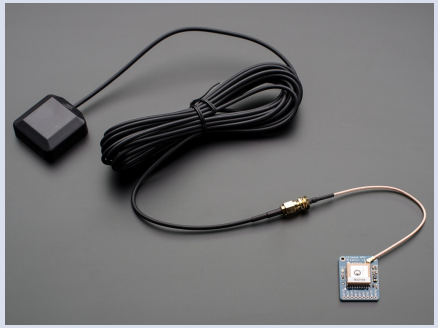


Adafruit GPS + antena
Sincronização dos relógios

CuBox Pro
Testes de vazão c/ TCP e UDP

Hardware de 2a Geração (2014) - Custo total aproximado ~R$ 1.200,00

Blue Appliance 847
Testes vazão e latência

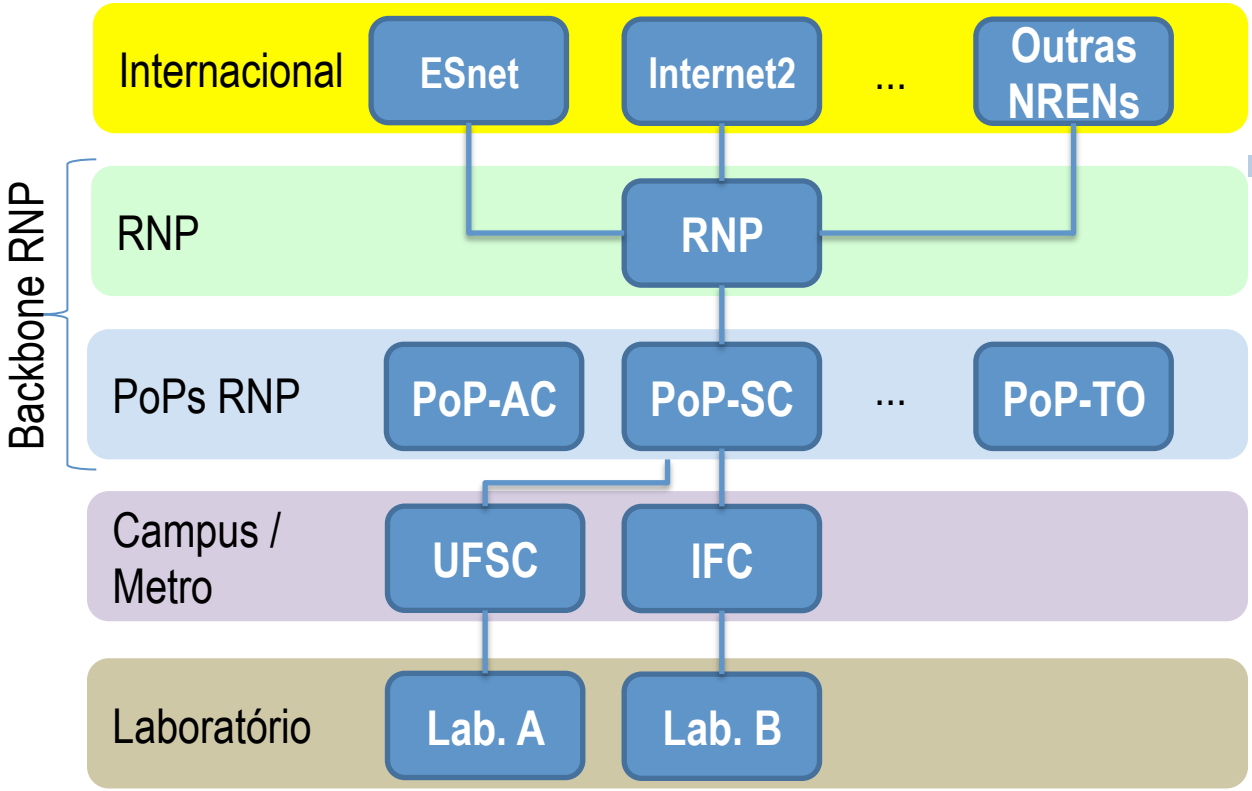


Adafruit GPS + antena
Sincronização dos relógios

Redução de 40% no custo total entre os kits da 1ª e 2ª versão

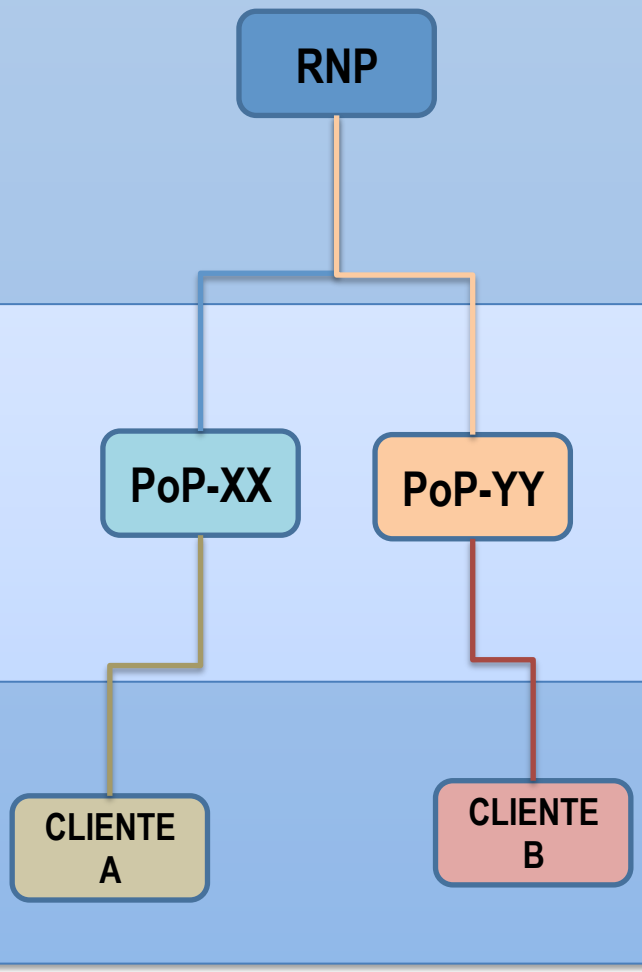
1. **Raspberry Pi** http://www.raspberrypi.org/
2. **CuBox** http://http://cubox-i.com/
3. **Componentes MonIPÊ** http://goo.gl/rNEFWO

MonIPÊ: arquitetura lógica em domínios



Domínio	Medições (em ou entre)
RNP	PoPs
PoP	PoP e instituição cliente diretamente conectada
Metro	PoP e Rede Metro
Campus	Campus e laboratórios de pesquisa ou outros campi
Laboratório	Dentro do campus
Internacional	Backbone RNP e outras NRENs

MonIPÊ: Componentes da Arquitetura

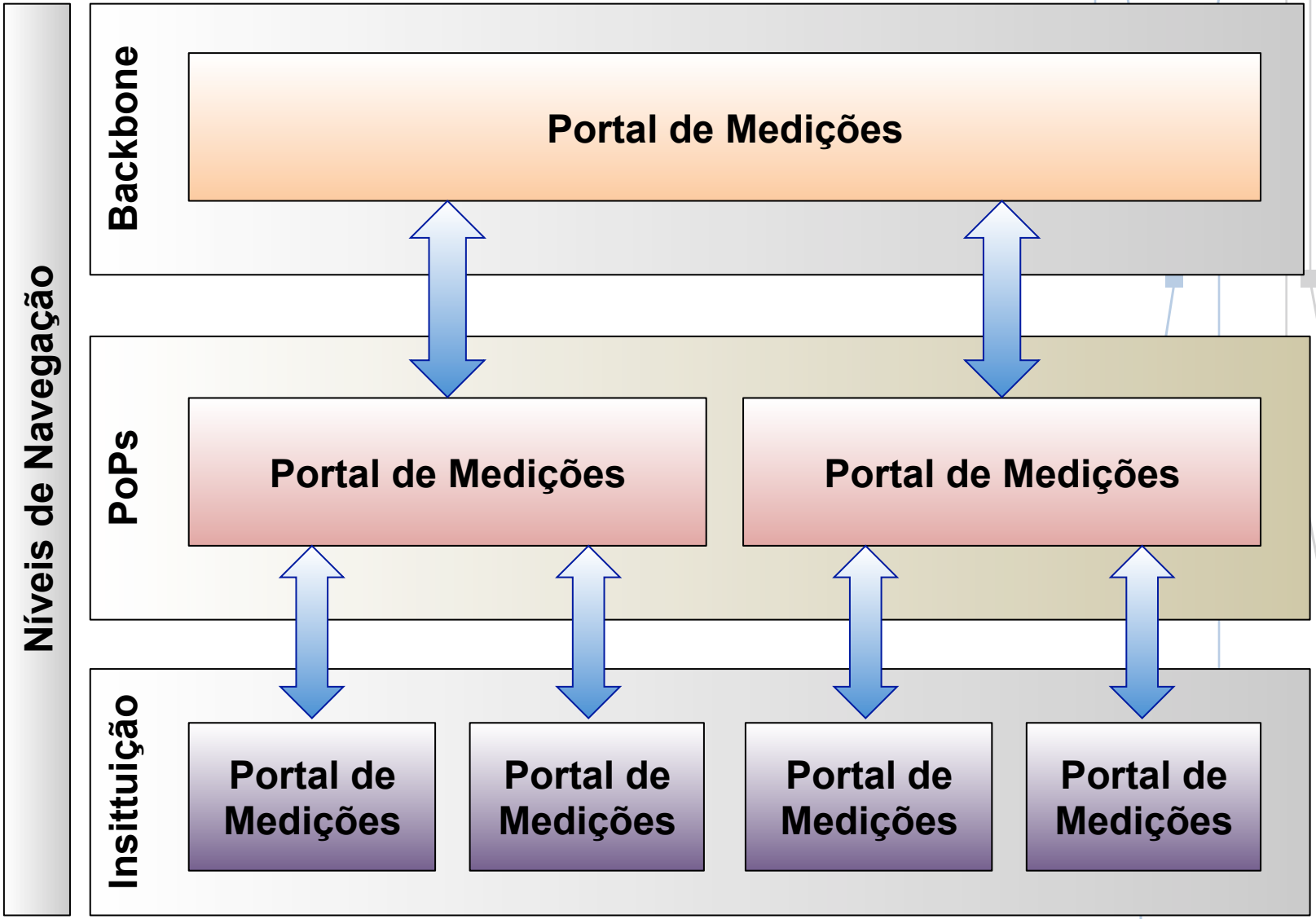
HIERARQUIA	COMPONENTES
	Domínio RNP • Portal de Medições • Módulo Agendamento • Módulo Sob-demanda • perfSONAR PHP-MA-SQL • perfSONAR Lookup Service • perfSONAR CLMP
	Domínio PoP • Portal de Medições • Módulo Agendamento • Módulo Sob-demanda • Antena GPS • perfSONAR PHP-MA-SQL • perfSONAR CLMP
	Domínio Instituição • Portal de Medições • Módulo Sob-demanda • Antena GPS de baixo custo • perfSONAR CLMP

MonIPÊ: Portal

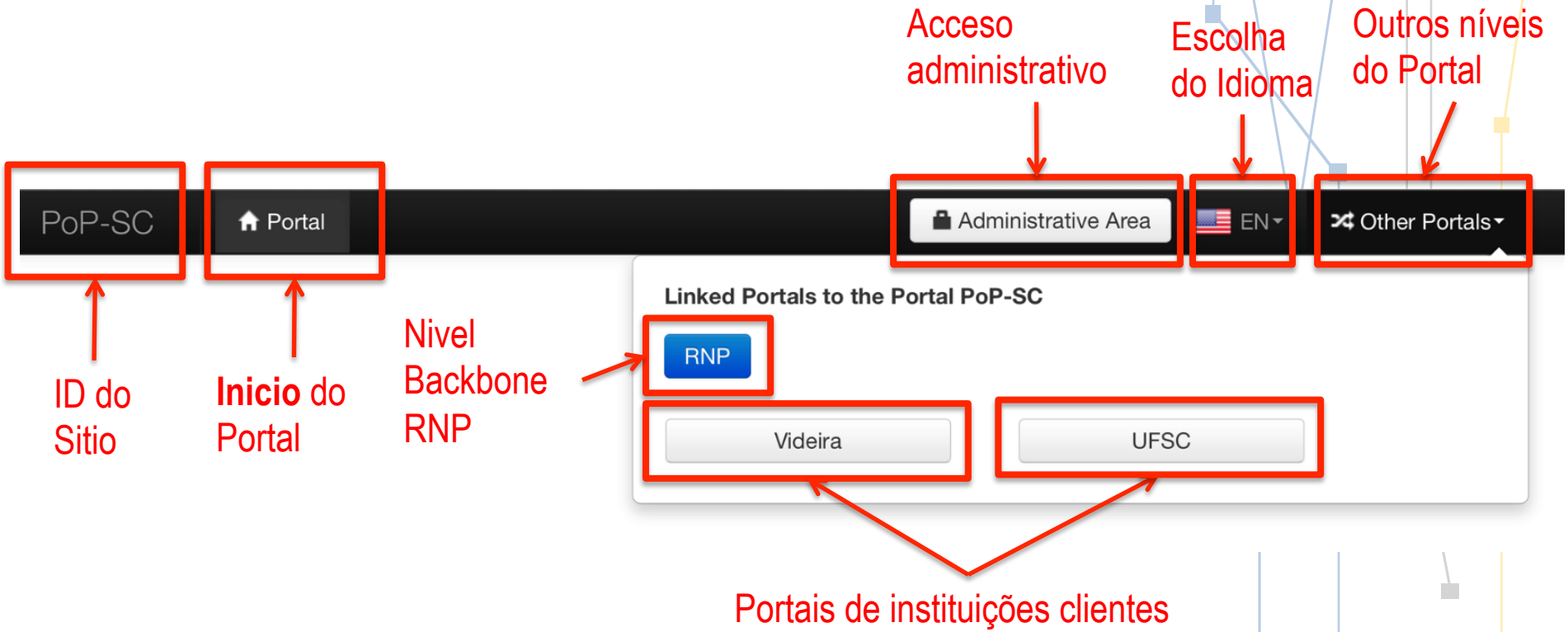
- **Interface gráfica web**
  - Configurações, gerenciamento, navegação multinível (domínios e *hosts*) e *feedback* de usuários (falhas/*bugs*, sugestões, críticas)
- **Agendamentos de testes:** sob demanda, temporários e periódicos
- **Armazenamento e recuperação** de medições
- **Visões:** pública e acesso restrito por senha (em desenv.: federado)
- **Integração** ao projeto perfSONAR (*Global Lookup Service* - GLS)
  - **Ferramentas:** ping, owamp, bwctl, traceroute* e ndt*
- **Testes** para MPs registrados no GLS
- **Atualizações:** manual ou automática

*** Em desenvolvimento**

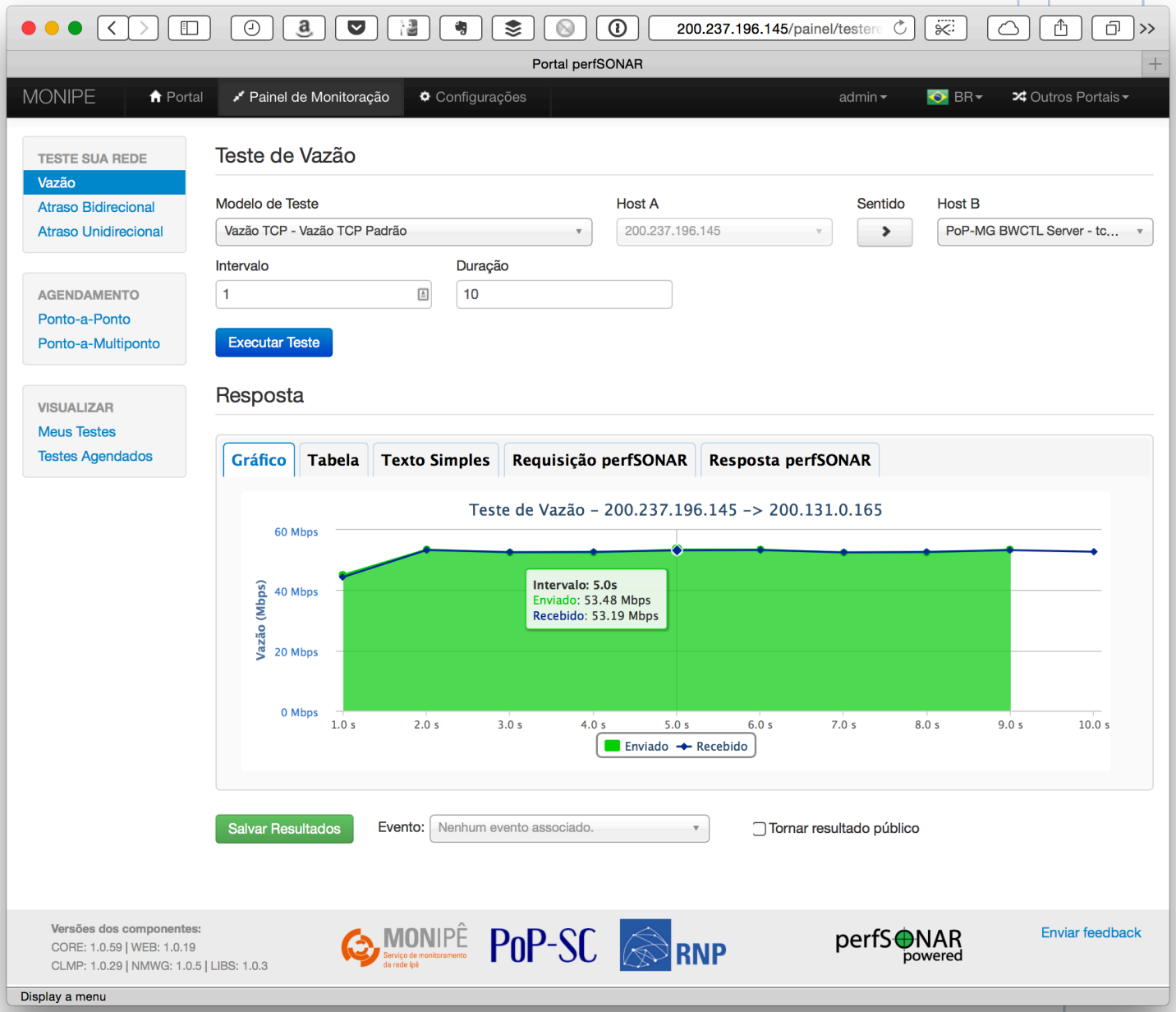
Portal web: navegação hierárquica entre os vários pontos de medições



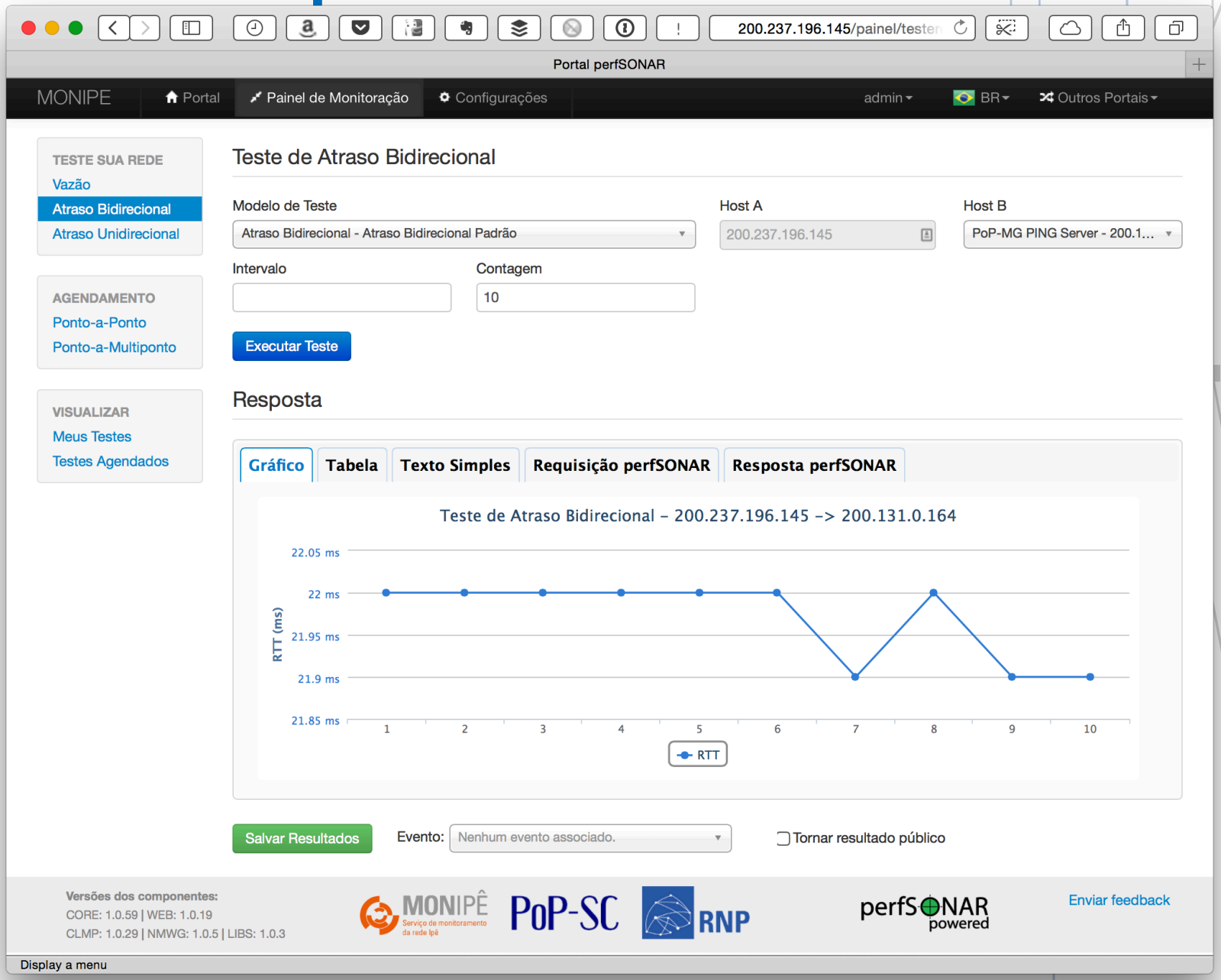
Interface do portal web



Interface do portal web: teste de vazão



Interface do portal web: teste de latência

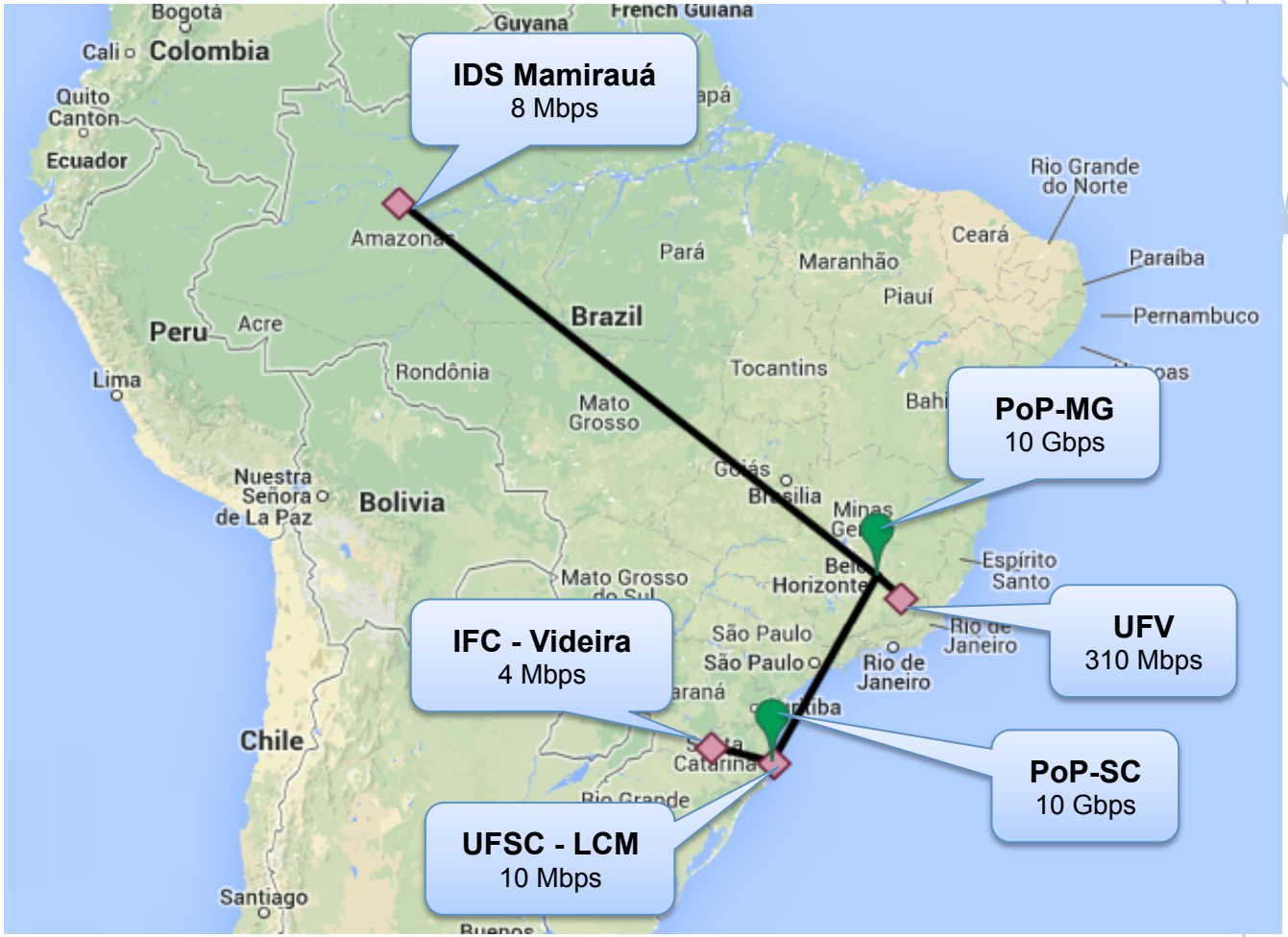


MonIPÊ: Realizações

- Protótipos construídos e testados:
  - MPs para medições até 1Gbps e até 10Gbps (2013-2014)
- Piloto realizado com 4 instituições (2013)
- Serviço em implantação no backbone RNP
- Prevista instalação em:
  - MPs VMs (em todos os PoPs)
  - MPs 10G (em 11 PoPs)
  - Kits baixo custo em ~100 instituições clientes da RNP
- Transição do Serviço Experimental MonIPÊ para produção, atendendo clientes da RNP
- Interoperabilidade com perfSONAR de outras redes

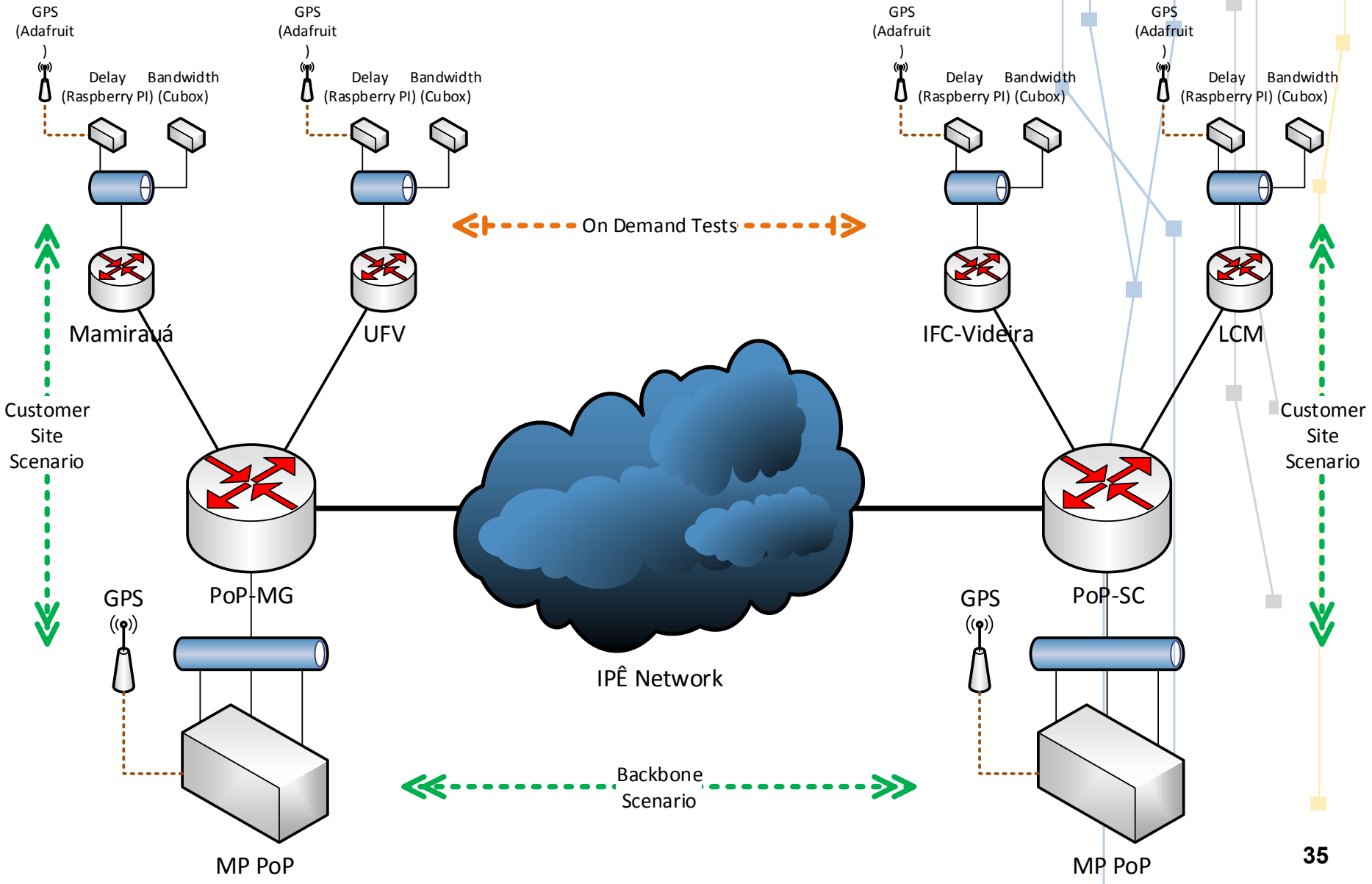
Piloto MonIPÊ (Nov/2013)

- **PoPs:** SC e MG (●)
- **Clientes:** Mamirauá, UFG, IFC-Videira e LCM/UFSC (◆)



Fonte: RNP Projeto MonIPÊ – Piloto 2013: http://goo.gl/gAJfMI

MonIPÊ: infraestrutura do Piloto (Nov/2013)



Evolução e Futuro: apoio do CT-Mon

CT-Mon: Comitê Técnico em Monitoramento de Redes

EQUIPE

- Coordenador Artur Ziviani (LNCC)
- Secretário: Alex Moura (RNP)
- Membros da comunidade acadêmica
- Acompanha a evolução do perfSONAR e em medições
- Apoia a RNP na evolução do Serviço MonIPÊ
- Colabora com o esforço de padronização do perfSONAR e em nível nacional e internacional
- Reuniões presenciais e remotas
- Apresentações de novos desenvolvimentos
- Tarefas específicas atribuídas a membros do comitê e/ou grupos de estudo/testes
- Chamadas temáticas de P&D de curta duração

TEMAS

Soluções e tecnologias para monitoramento:

- Medições ativas e passivas: 1G, 10G, 100Gbps+
- Redes sem fio
- Metodologias, técnicas e ferramentas
- Grandes eventos: Copa do Mundo, Olimpíadas
- Transferências de grandes volumes de dados
- Monitoração nas camadas abaixo da camada 3
- Geração de alertas (previsão de problemas)
- Armazenamento, compactação recuperação e compartilhamento de dados históricos
- Evolução do tráfego na RNP, QoE, DPI etc…

Equipe

MONIPÊ Serviço de monitoramento da rede Ipê

perfSONAR powered

PoP-SC RNP

COORDENAÇÃO

RNP - Diretoria de Pesquisa e Desenvolvimento

Direção	Michael Stanton
Direção Adjunta	Iara Machado
Gerência	Alex Soares de Moura
Coordenação	Marcos Schwarz
Coordenação	Fausto Vetter

Desenvolvimento

PoP-SC - Ponto de Presença de Santa Catarina (UFSC)

Coordenação

Coordenação Administrativa do Desenvolvimento	Edison Melo
Coordenação Técnica do Desenvolvimento	Murilo Vetter

Infraestrutura - Desenvolvimento de Hardware

Desenvolvimento de Hardware e Infraestrutura	Rodrigo Pescador
Desenvolvimento de Hardware e Infraestrutura	Guilherme Rhoden
Desenvolvimento de Hardware e Infraestrutura	Rodrigo Gonçalves
Desenvolvimento de Software Web	Paulo Brandtner
Desenvolvimento de Software Web	Luis Fernando Cordeiro

BOLSISTAS

Desenvolvimento de Software	Leonardo Schlüter Leite
Desenvolvimento de Software	Kádio Francisco Miguel Colzani

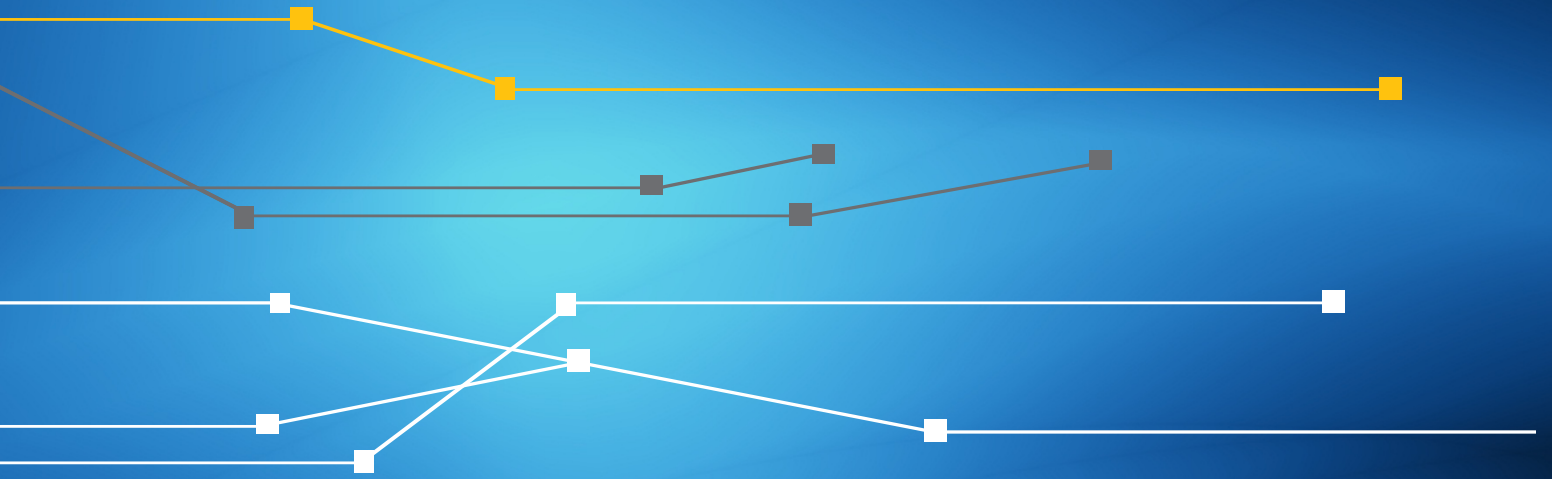
Agradecimentos

PoP-SC

perfSONAR

UFSC

UFBA



Obrigado!

Alex Soares de Moura
alex.moura@rnp.br

Rede Nacional de Ensino e Pesquisa – RNP
Ponto de Presença da RNP em Santa Catarina - PoP-SC
Universidade Federal de Santa Catarina – UFSC
Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação - SeTIC/UFSC

RNP

Ministério da **Cultura**

Ministério da **Saúde**

Ministério da **Educação**

Ministério da **Ciência, Tecnologia e Inovação**