LIVE Demo: Transmissão de dados a 100Gbps

Introdução

07-2019

IMPLANTAÇÃO DOS 100G NO NORDESTE

Disponibilidade dos trechos BA-PE PE-PB PB-RN RN-CE

01-2020

PLANEJAMENTO DA DEMONSTRAÇÃO 100G

Formação da equipe e organização das atividades

11-2020

FECHAMENTO DO ANEL NORDESTE

Implantação do trecho CE-BA

12-2020

APRESENTAÇÃO NO WRNP

Apresentação da plataforma no WRNP

Introdução

- Tráfego de dados efetivo nos enlaces menor que **7Gbps**
  - Ainda menor durante a pandemia
  - Necessidade de validar a capacidade também para efeito prático

Iniciativa

- Demonstrar a capacidade efetiva do backbone, nos enlaces ativos do Nordeste, que na época formavam o maior conjunto de PoPs ativos a 100Gbps
- Objetivo é construir uma plataforma de demonstração simplificada (similar ao *Speedtest*)
- Grupo de trabalho entre a DPD/RNP e os PoPs (BA, CE, PB, PE, RN) para construção da solução

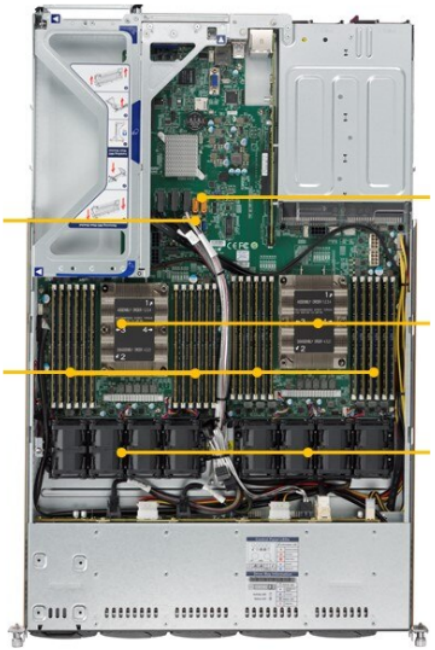
Iniciativa

- Equipamentos alocados para a construção de nós DTN (Projeto IDS da DPD/RNP) disponíveis
  - Alguns já instalados nos PoPs, porém com interfaces de rede para trafegar no máximo 10Gbps
- Interfaces para conexão a 100Gbps disponíveis nos roteadores do backbone

Site da Demonstração

https://demo100g.rnp.br

Equipamentos



* Imagens: *Servidor do projeto IDS*

Equipamentos

- 2 x Sockets (Intel scalable processors 1ª e 2ª geração)
- 24 x Slots de memória DDR4 2966MHz, 12 para cada CPU
- **2 x Slots de expansão PCIe 3.0 x16 perfil alto (FP)**
- **2 x Slots de PCIe x8 perfil baixo (LP)**
- 2 x Interfaces 10GbE ótica (SFP+)
- 2 x Interfaces 1GbE cobre
- 10 x HDDs 2,5" SAS3/SATA

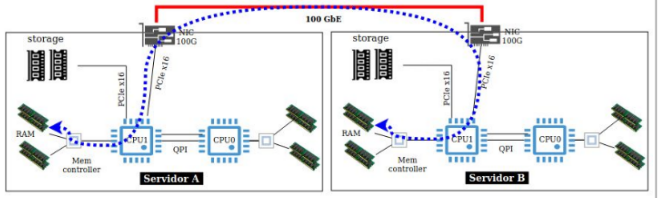
- Embora o servidor não tenha sido especificado para operar com interfaces 100 GbE, tinha potencial para ser aproveitado:
  - Possibilidade de inserção de placas 100Gbps nos slots *PCIe*

Proposição

- Aquisição e logística de placas e cabeamentos necessários
- Validação da solução de transferência de dados (backend)
- Avaliação da solução de visualização da transferência (frontend)
  - Coletor dos dados (equipamentos e aplicações)
  - Gráficos de uso dos recursos (métricas de consumo)
  - Apresentação do tráfego real (passante)
  - Medidor de banda corrente (estilo velocímetro)

Proposição

- Backend avaliado em laboratório pela DPD/RNP

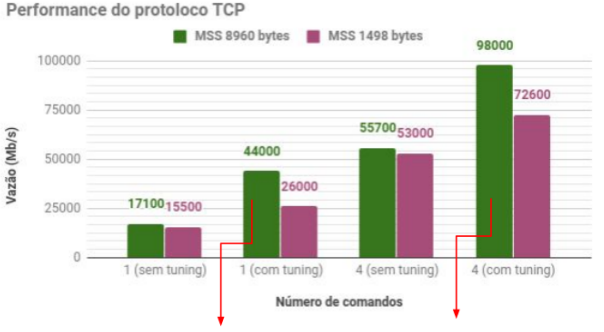
Descrição	Topologia
Objetivo: Checar a performance de transferência na rede entre os servidores conectados a 100G, usando protocolo TCP; O teste é feito da RAM do servidor A para RAM do servidor B e vice-versa;	

Servidor:
numactl --preferred=<numa> --cpunodebind=<cpu> iperf3 -s -D -p <n_porta>

Cliente:
numactl --preferred=<numa> --cpunodebind=<cpu> iperf3 -p <n_porta> -c <IP_cliente> --parallel 1 -t 65 --omit 5 --set-mss [8960/1498]

Para múltiplas threads, executar vários dos comandos acima em paralelo, usando um número de porta diferente para cada comando.

Notas:
1. Nos testes com o 1 fluxo, o uso do Core no receptor chegou a 98% indicando que para alcançar maior vazão é necessário utilizar outros Cores simultaneamente;
2. Aumentar o número de fluxos TCP paralelos no iperf3 (-P) não mostrou ganhos significativos. E '-P=4', por exemplo, não significa que 4 threads (Cores) diferentes serão utilizados ao mesmo tempo.



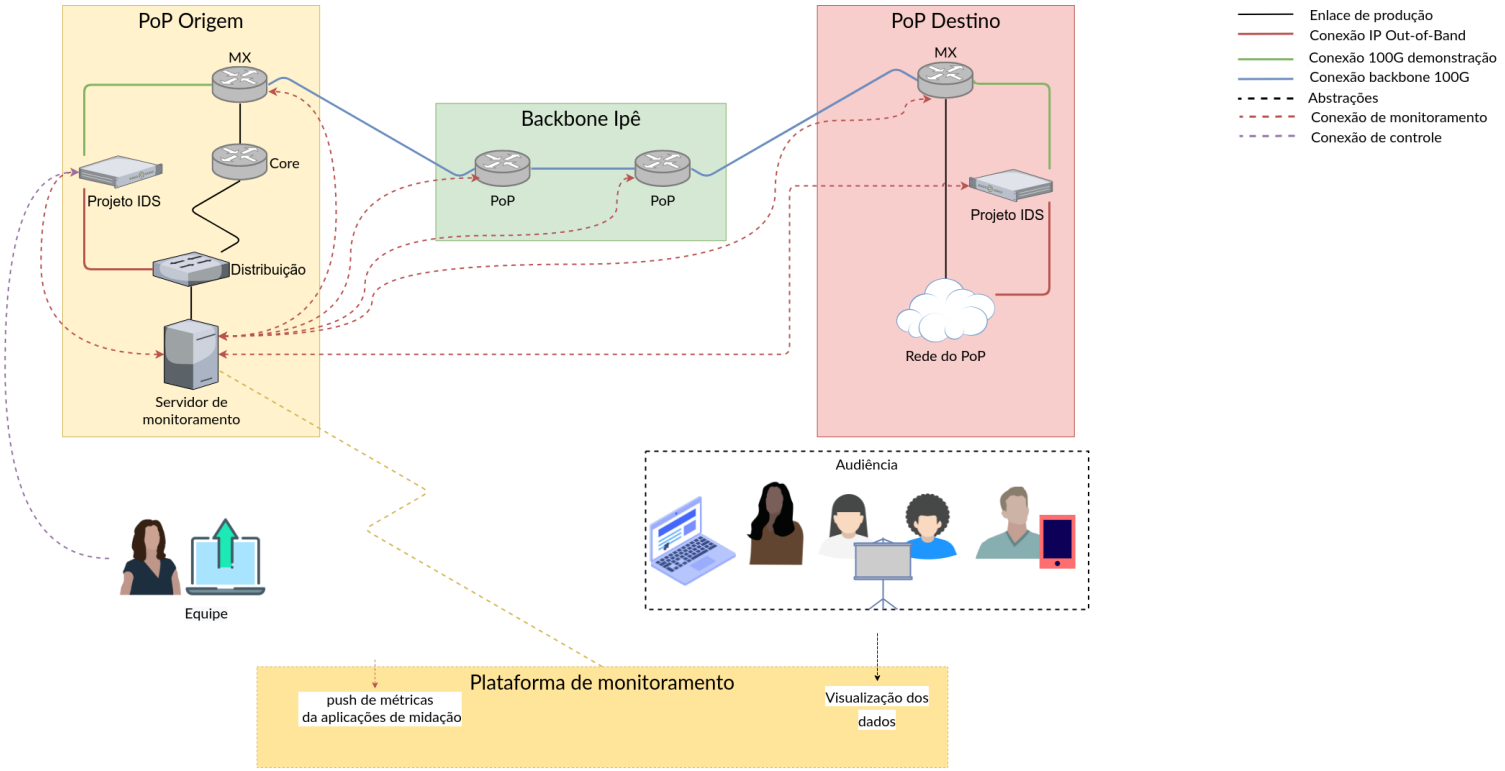
Número de retransmissões = 0

* Imagem: *Avaliação P2P dos servidores IDS para transmissões 100Gbps*

Proposição

- Ferramentas e modelos iniciais de embasamento
  - Speedtest Custom
  - Speedtest CLI
  - LibreSpeed
  - OpenSpeedTest
  - NDT

Proposição



* Imagem: *Arquitetura inicial*

Time-Series DataBase (TSDB)

Time-Series DataBase (TSDB)

- Permite formas eficientes de recuperação e inserção dos dados relacionados às métricas monitoradas
- Evita consumo desnecessário de espaço em disco e armazena métricas por longo períodos de tempo
- Na escolha de uma ferramenta de monitoramento de métricas temporais é necessário considerar o ecossistema da ferramenta escolhida:
  - Coletores, encaminhadores (proxies), monitoradores, visualizadores (front-end) e ferramenta de armazenamento (back-end)

Time-Series DataBase (TSDB)

- **RRDTool**: Dados são mantidos em forma circular
- **Graphite**: Possui o TSDB com base de dados de tamanho fixo
- **InfluxDB**: Data store de alto desempenho
- **OpenTSDB**: Distribuído, escalável (arquitetura de alta disponibilidade)
- **TimescaleDB**: Implementado como extensão do PostgreSQL, suportando as mesmas operações e queries SQL

Prometheus

an open-source systems monitoring and alerting toolkit

Principais funcionalidades

- Modelo multidimensional de dados: séries de dados identificadas por métricas e pares de chave/valor
- Linguagem de consulta flexível (**PromQL**)
  - Suporte a operadores lógicos, aritméticos e agregadores (sum, min, avg, etc)
- Não depende de armazenamento distribuído; nós de um único servidor são autônomos
- As coletas de séries temporais são realizadas via HTTP Pull (**exporter**)

Componentes

- Servidor de coleta (**scrapes**) e banco de dados de séries temporais (**Prometheus server**)
- Bibliotecas de cliente (**client libaries**) para programação
- **push gateway** para suportar jobs de curta duração
- Exportadores (**exporters**) para diversos serviços: SNMP, StatsD, Graphite, etc
- O **alertmanager** para disparar alertas
- Outras ferramentas de suporte

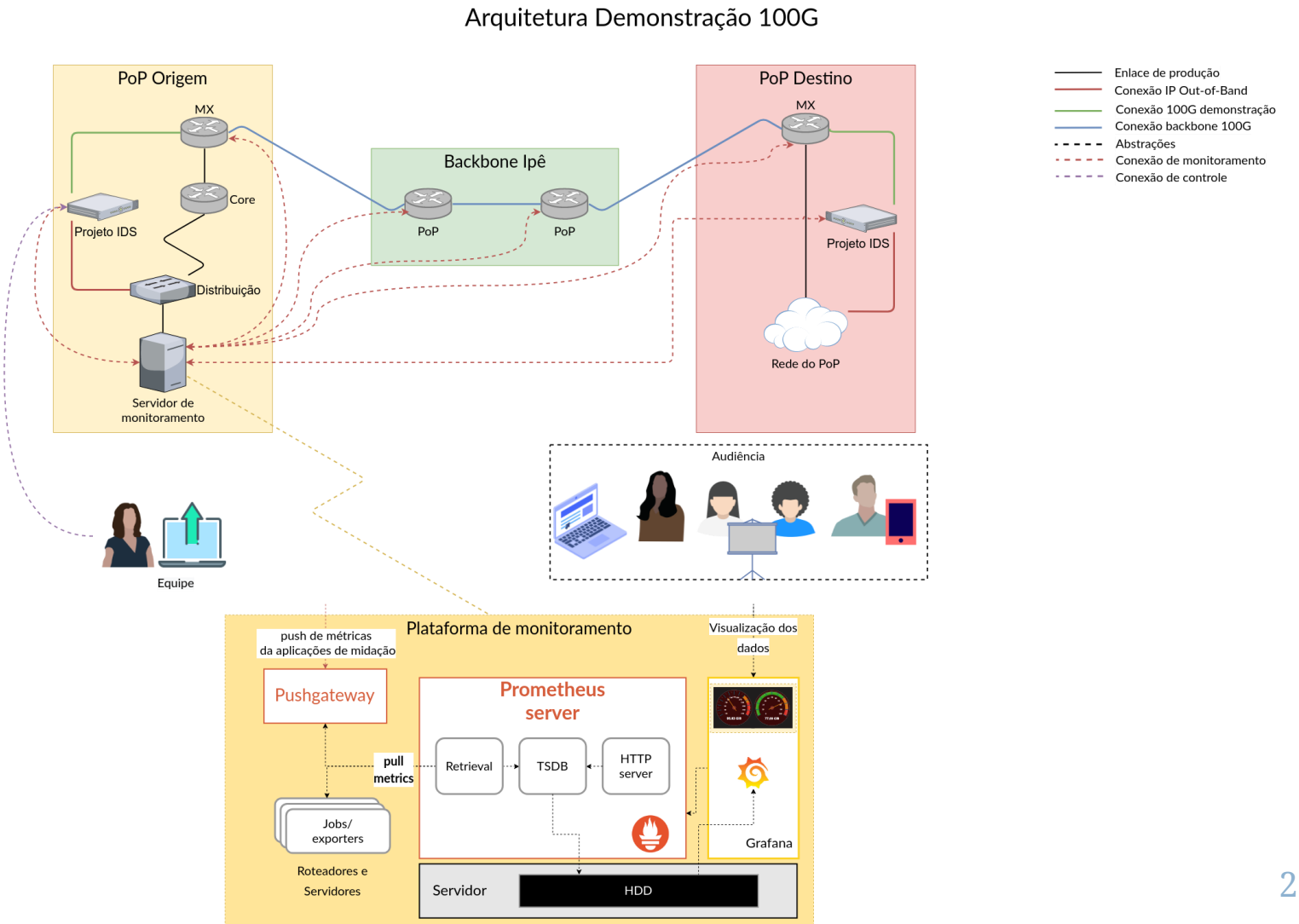
Grafana

an open observability platform

Grafana

- Plataforma para monitoramento e observabilidade (monitoring e observability)
- Visualização e suporte para diversas fontes de dados (data sources), incluindo **Prometheus**
- Realização de buscas, visualização e emissão de alertas
- Criação de múltiplos dashboards e disponibilização de um conjunto de templates
- Criação de gráficos com combinações de múltiplos datasets (mixed data sources)

Proposição


Arquitetura Demonstração 100G
PoP Origem
MX
Core
Projeto IDS
Distribuição
Servidor de monitoramento
Backbone Ipê
PoP
PoP
PoP Destino
MX
Projeto IDS
Rede do PoP
Enlace de produção
Conexão IP Out-of-Band
Conexão 100G demonstração
Conexão backbone 100G
Abstrações
Conexão de monitoramento
Conexão de controle
Audiência
Equipe
Plataforma de monitoramento
push de métricas da aplicações de midação
Pushgateway
pull metrics
Prometheus server
Retrieval
TSDB
HTTP server
Jobs/ exporters
Roteadores e Servidores
Servidor
HDD
Visualização dos dados
Grafana

POC

Início dos testes da proposta

POC

- Criação do ambiente de teste
- Construção de container similar ao ambiente de produção
  - Instalação e configuração do Prometheus
  - *exporters*: Prometheus, Node e SNMP
- Configuração de equipamentos para coleta (Servidor e equipamento de rede)

POC

- Coleta e medição dos dados em intervalos de *500ms*

```
2818629252752 @1594379981.787
2818629252752 @1594379982.287
2818629252752 @1594379982.787
2818629252752 @1594379983.287
2818629252752 @1594379983.787
2818751807315 @1594379984.287
2818751807315 @1594379984.787
2818751807315 @1594379985.286
2818751807315 @1594379985.787
2818751807315 @1594379986.787
2818751807315 @1594379987.286
```

* Imagem: *Consulta ao Prometheus - SNMP*

POC

- Coleta e medição dos dados em intervalos de *500ms*

```
1342983084145 @1594382373.375
1342983084145 @1594382373.875
1342983084145 @1594382374.375
1342985047823 @1594382374.875
1342985047823 @1594382375.375
1342985047823 @1594382375.875
1342985047823 @1594382376.375
1342985047823 @1594382376.875
1342985047823 @1594382377.375
1342990433793 @1594382377.875
1342990433793 @1594382378.375
1342990433793 @1594382378.875
```

* Imagem: *Consulta ao Prometheus - SNMP - Servidor*

POC

- Métricas coletadas via SNMP
  - Limitado a certos recursos disponibilizados pela implementação do fabricante
  - **Necessita de rotinas internas de agregação de informação no plano de controle**

POC da proposta

- Coleta e medição dos dados em intervalos de *200ms*
- Apenas para coleta nos servidores (endpoints)

```
17226998229 @1594382942.006
17227002296 @1594382942.506
17227007414 @1594382943.006
17227011481 @1594382943.506
17227015548 @1594382944.006
17227019747 @1594382944.506
17227023814 @1594382945.006
17227028932 @1594382945.506
17227032999 @1594382946.006
17227038117 @1594382946.506
17227042184 @1594382947.006
17227047302 @1594382947.506
```

* Imagem: *Consulta ao Prometheus - Node - Servidor*

Telemetria

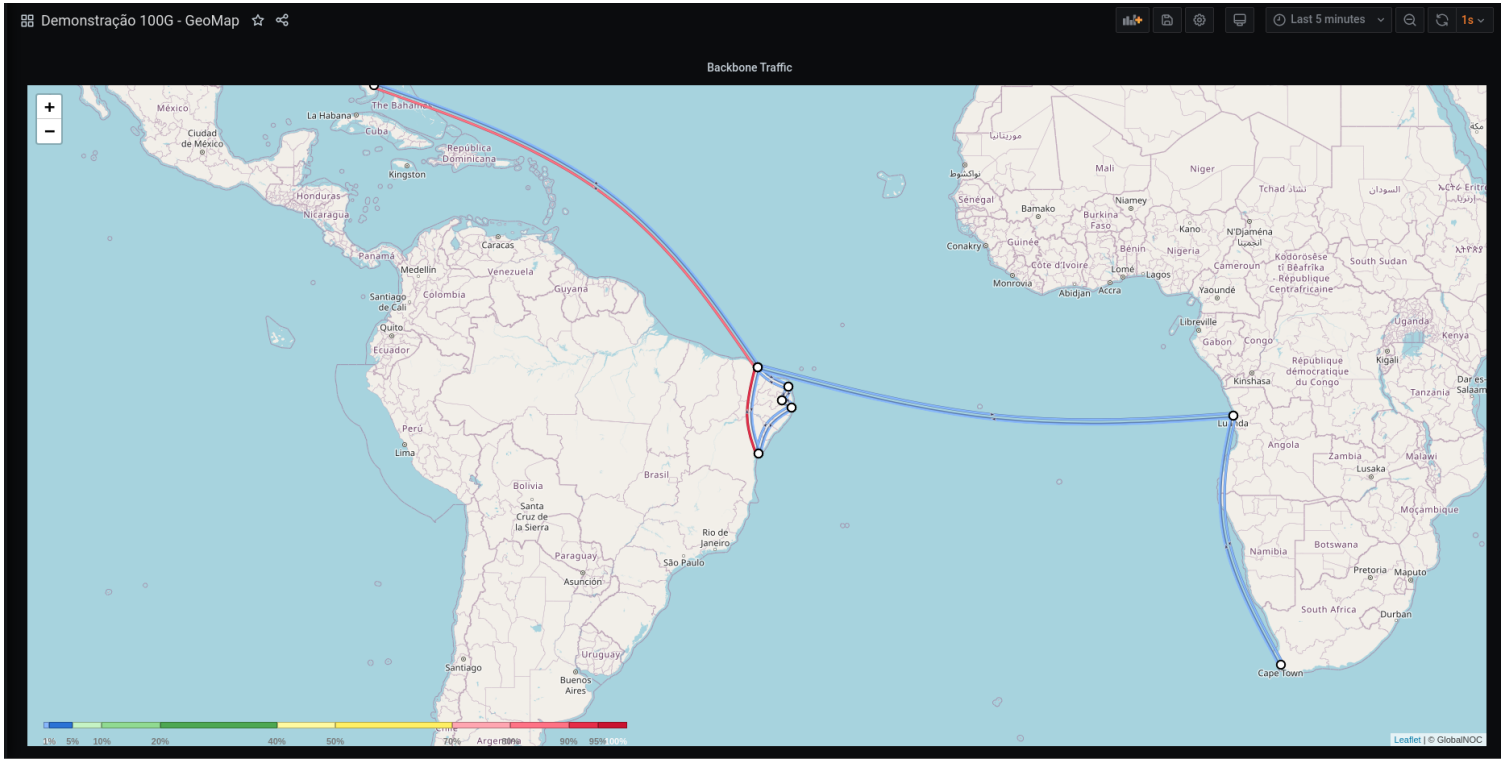
Telemetria

- Permite o sensoriamento através de uma interface mais universal e granular
  - Eventos interfaces, *buffer*, etc.
  - Adoção de modelo *PUSH*
- Diversas propostas
  - In-band Network Telemetry (INT)
  - *Junos Telemetry Interface (JTI)*
  - Cisco Model-Driven Telemetry (Streaming Telemetry)

Apresentação

* Imagem: *Painel do Prometheus - Servidor*

Apresentação

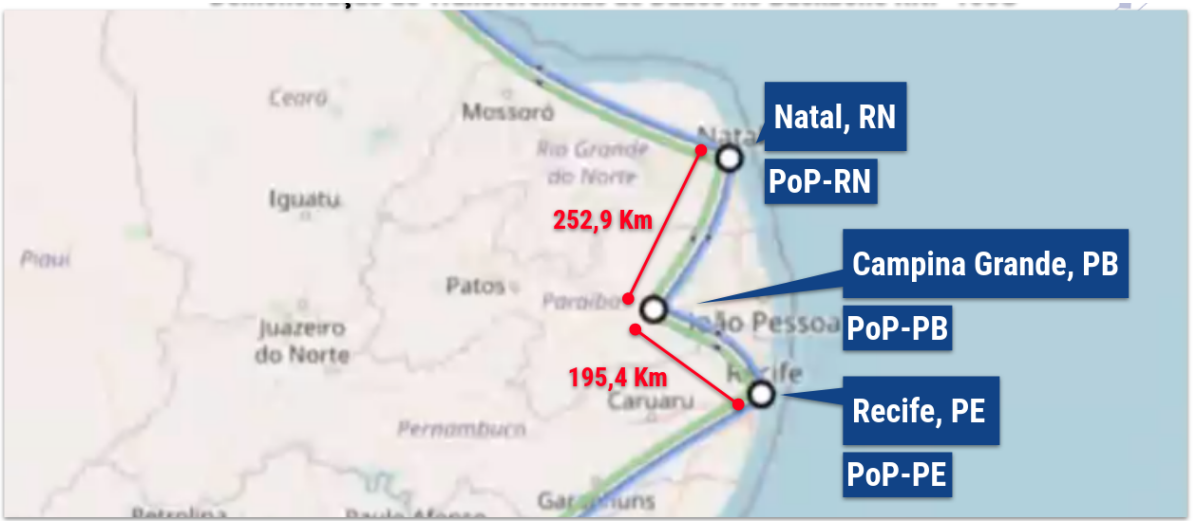


* Imagem: *Painel do Prometheus - GeoMap*

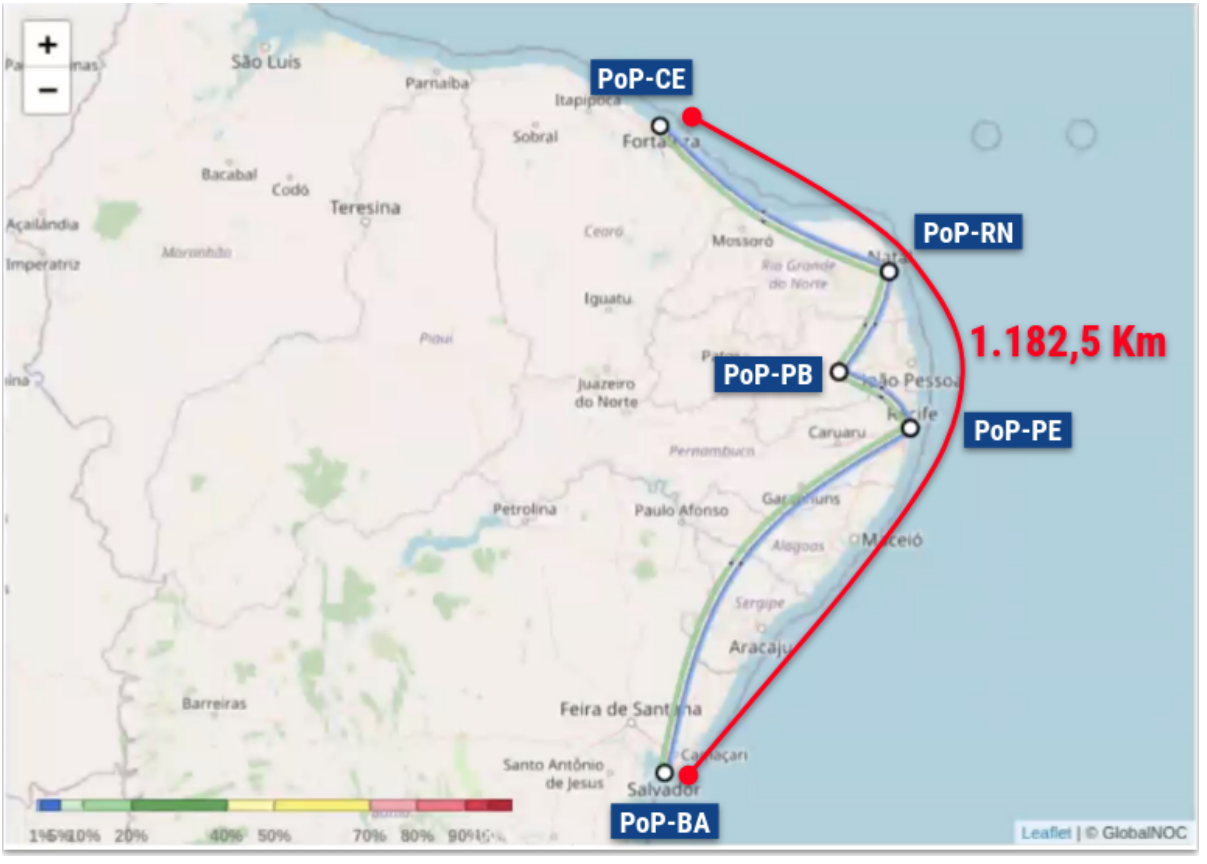
Topologia


Recife
Caruaru
Pernambuco
Petrolina
Paulo Afonso
Maceió
Alagoas
809,4 Km
Sergipe
Aracaju
Feira de Santana
Santo Antônio de Jesus
Salvador
Recife, PE
PoP-PE
Salvador, BA
PoP-BA

Topologia


Ceará
Mossoró
Rio Grande do Norte
Iguatu
Piauí
Patos
Paraíba
João Pessoa
Juazeiro do Norte
Recife
Caruaru
Pernambuco
Natal, RN
PoP-RN
252,9 Km
Campina Grande, PB
PoP-PB
195,4 Km
Recife, PE
PoP-PE

Topologia


PoP-CE
PoP-RN
PoP-PB
PoP-PE
PoP-BA
1.182,5 Km
São Luís
Parnaíba
Sobral
Bacabal
Codó
Teresina
Açailândia
Imperatriz
Iguatu
Juazeiro do Norte
Petrolina
Paulo Afonso
Caruaru
Mossoró
Barreiras
Feira de Santana
Santo Antônio de Jesus
Salvador
Aracaju
Leaflet | © GlobalNOC

Demonstração

Agradecimentos

Referências

1. **Monitoring Distributed Systems**: https://landing.google.com/sre/sre-book/chapters/monitoring-distributed-systems/
2. BADER, Andreas. **Comparison of Time Series Databases**. 2016. Tese de Doutorado. Diploma Thesis, Institute of Parallel and Distributed Systems, University of Stuttgart.
3. PETRE, Ionut et al. **A Time-Series Database Analysis Based on a Multi-attribute Maturity Model**. Studies in Informatics and Control, v. 28, n. 2, p. 177-188, 2019.
4. BADER, Andreas; KOPP, Oliver; FALKENTHAL, Michael. **Survey and Comparison of Open Source Time Series Databases**. In: BTW (Workshops). 2017. p. 249-268.
5. **Streaming Telemetry**: https://developer.cisco.com/docs/ios-xe/#!streaming-telemetry-quick-start-guide/streaming-telemetry
6. **In-band Network Telemetry (INT**: https://p4.org/assets/INT-current-spec.pdf
7. **Junos Telemetry Interface User Guide**: https://www.juniper.net/documentation/en_US/junos/information-products/pathway-pages/junos-telemetry-interface/junos-telemetry-interface.pdf

Referências (Complementar)

- **DPDtechtalk 1: Arquitutra de servidores para operar a 100GbE**: DPD/RNP, Março/2020.
- **The USE Method**: http://www.brendangregg.com/usemethod.html
- **The RED Method**: key metrics for microservices architecture: https://www.weave.works/blog/the-red-method-key-metrics-for-microservices-architecture/
- **Prometheus + Grafana + Gauge**: https://wiki.openvz.org/SNMPD_in_container
- **SNMP Exporter**: https://labs.consol.de/omd/howtos/prometheus_snmp_exporter/
- **snmp_exporter - SNMP Exporter for Prometheus**: https://www.diycode.cc/projects/prometheus/snmp_exporter
- **Prometheus: Prometheus monitoring switch (snmp)**: https://programmer.group/prometheus-prometheus-monitoring-switch-snmp.html
- **Prometheus: snmp_exporter and OpenBSD**: https://yetiops.net/posts/openbsd-snmp-exporter/
- **JTI Plug-ins for Open Source Data Collectors**: https://www.juniper.net/documentation/en_US/junos/topics/concept/jti-opensource-plugins.html