

Implantação e Avaliação de Desempenho de Protocolos de Transmissão Multicast Confiável na RNP

**Valter Roesler, Marinho P. Barcellos
Evandro C. Dall’Agnol, Giovani Facchini
Gustavo Bervian Brand, Renato Costa,
Tasso Gomes de Farias**

Universidade do Vale do Rio dos Sinos (UNISINOS)
Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
PRAV – Laboratório de Redes de Alta Velocidade

<http://prav.unisinos.br/gtmc>

Sumário

- Introdução: objetivos, conceitos, tecnologias e protocolos
- Resultados
 - Rede Local
 - Agregado Giga
 - RNP
- Interface com o Usuário
- Conclusões
- Próximos passos



Objetivos do Grupo

- Montar uma estrutura de multicast confiável na RNP, explorando o suporte nativo à multicast atualmente existente
- Avaliar experimentalmente um conjunto de implementações de protocolos de multicast confiável
- Definir metodologia para reavaliar a rede e os protocolos em caso de mudança de topologia



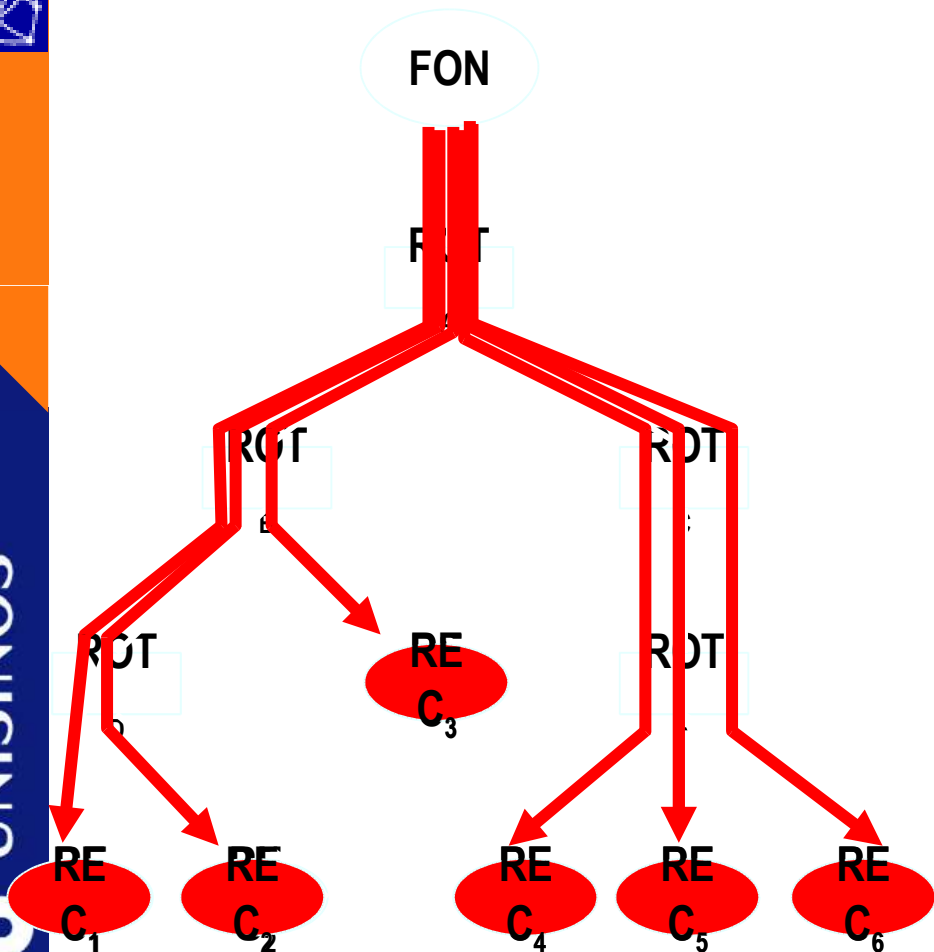
Conceitos

- IP Multicast: roteamento de pacotes pela rede, com replicação eficiente segundo uma árvore
- Multicast Confiável: funções de transporte, como confiabilidade, controle de sessão e congestionamento
- Exemplos de aplicações multicast: disseminação maciça de dados, transmissão multimídia 1xn, videoconferências e outras aplicações multi-participante interativas, incluindo jogos online

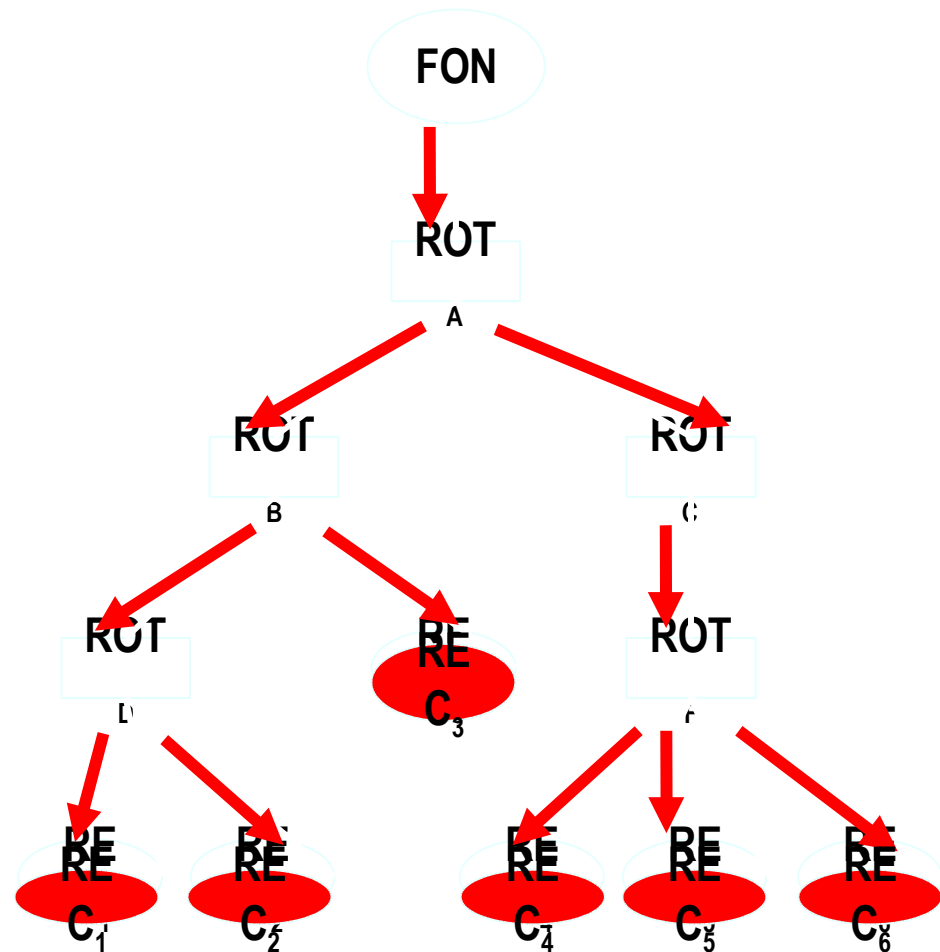


Multiponto sem e com Multicast

MULTI-UNICAST



MULTICAST





Exemplos de Aplicações multicast

- Transmissão maciça de dados (foco do GT)
- Aplicações tipo “push”
- Transmissões multimídia unidirecionais ao vivo (TV, rádio, palestras, conferências)
- Jogos interativos

Requisitos Típicos para Aplicações Multicast

- Latência de entrega
- *Jitter*
- *Skew*

Tempo-real (multimídia)

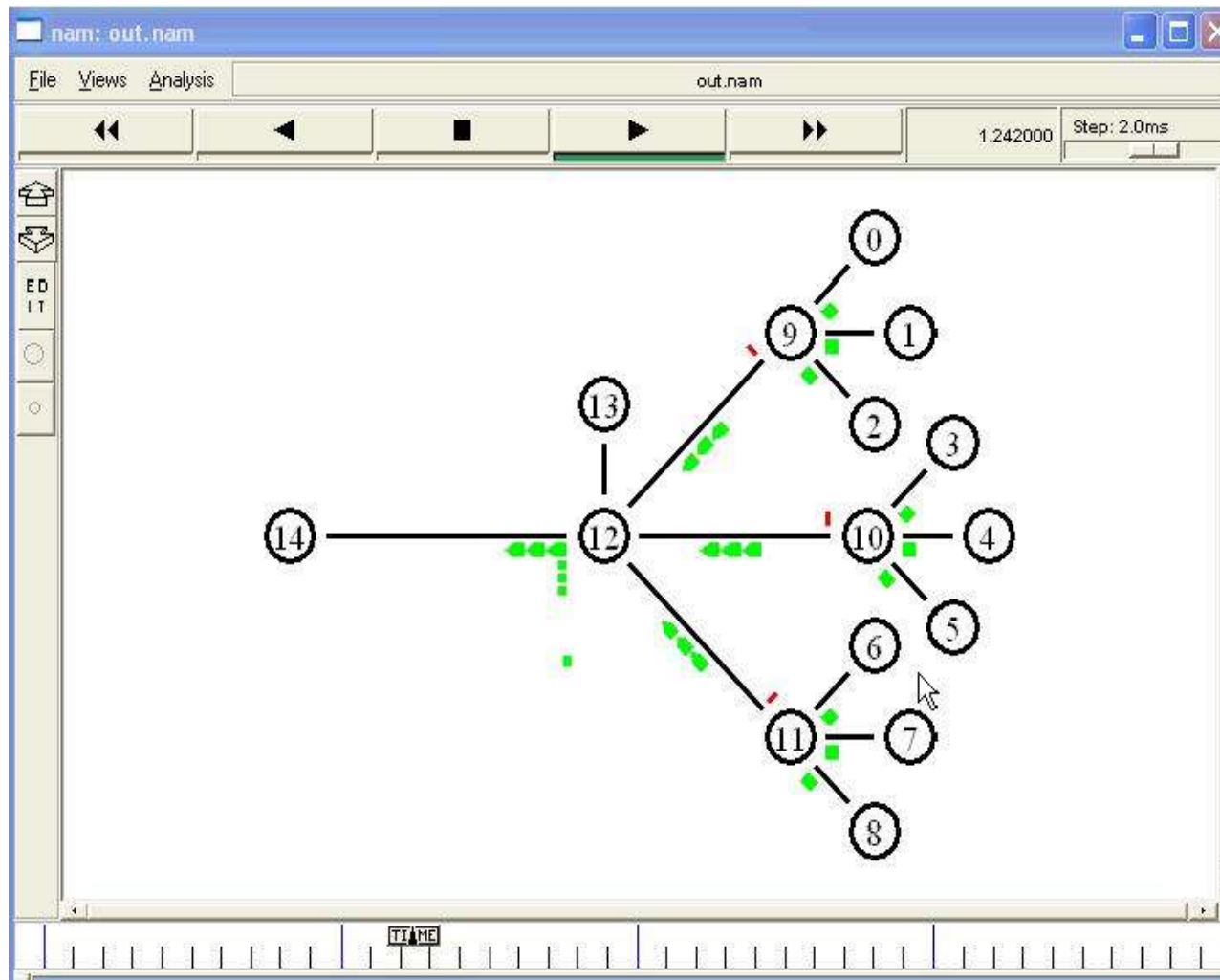
- Ordenamento de pacotes
- Confiabilidade

Confiável (arquivos)



Problemas do multicast confiável

- Implosão de ACK e NACK



Padronização no Multicast Confiável



- IETF: Working group on Reliable Multicast Transport
- RFC 3048 (Reliable Multicast Transport Building Blocks ...): padroniza sistemas de multicast confiável através de dois conceitos:
 - BB (*Building Blocks*): componentes comuns utilizados por vários protocolos, como mecanismos de FEC, confirmações negativas (NACKs), controle de congestionamento e segurança.
 - PI (*Protocol Instantiations*): são os protocolos propriamente ditos, que se utilizam dos BBs necessários e inserem sua lógica de funcionamento.

Famílias de protocolos definidas pelo IETF (RFC 3048)

- NACK-based Reliable Multicast Protocol (NORM)
 - Protocolos baseados em NACK
- Tree-based ACK
 - Protocolos baseados em ACK, porém com servidores intermediários a fim de evitar implosões de ACKs. OBS: foi obsoleto
- Asynchronous Layered Coding (ALC)
 - Protocolos baseados em FEC, não necessitando retransmissões
- Router Assist
 - Protocolos que necessitam código rodando nos roteadores intermediários. OBS: necessitam redes ativas ou cooperação do fabricante de roteadores, ou seja, atualmente não é viável praticamente



Protocolos Investigados

- Nack + FEC
 - MDP/NRL
 - NORM/NRL
 - NORM/INRIA
- Camadas + FEC
 - ALC/INRIA
 - ALC/Tampere UT
- FEC
 - DF
- ARQ
 - TCP-XM
 - MultiTCP
- Hierárquico
 - JRMS/Sun Microsystems



Recursos necessários nos POPs

- Acesso remoto a máquinas Linux nos POPs utilizados;
- Conectividade multicast
- Auxílio eventual do pessoal da RNP a fim de resolver problemas eventuais de mal funcionamento dos micros ou na rede



Preparação do ambiente 1

- Instalação de software no diretório home (/gt/gtmc) dos POPs que o GT iria utilizar
 - gcc 3.3.5
 - g++ 3.3.5
 - glibc 2.3.2
 - libstdc++ 3.3.5
 - python 2.3.4
 - java: JVM - SDK 1.4.2_07
 - permissão para uso do tcpdump via sudo



Preparação do ambiente 2

- Definição dos parâmetros de cada protocolo
 - Feito através de testes com taxas baixas de transferência
 - Duração de várias semanas, pois envolvia diversos testes com todos os protocolos



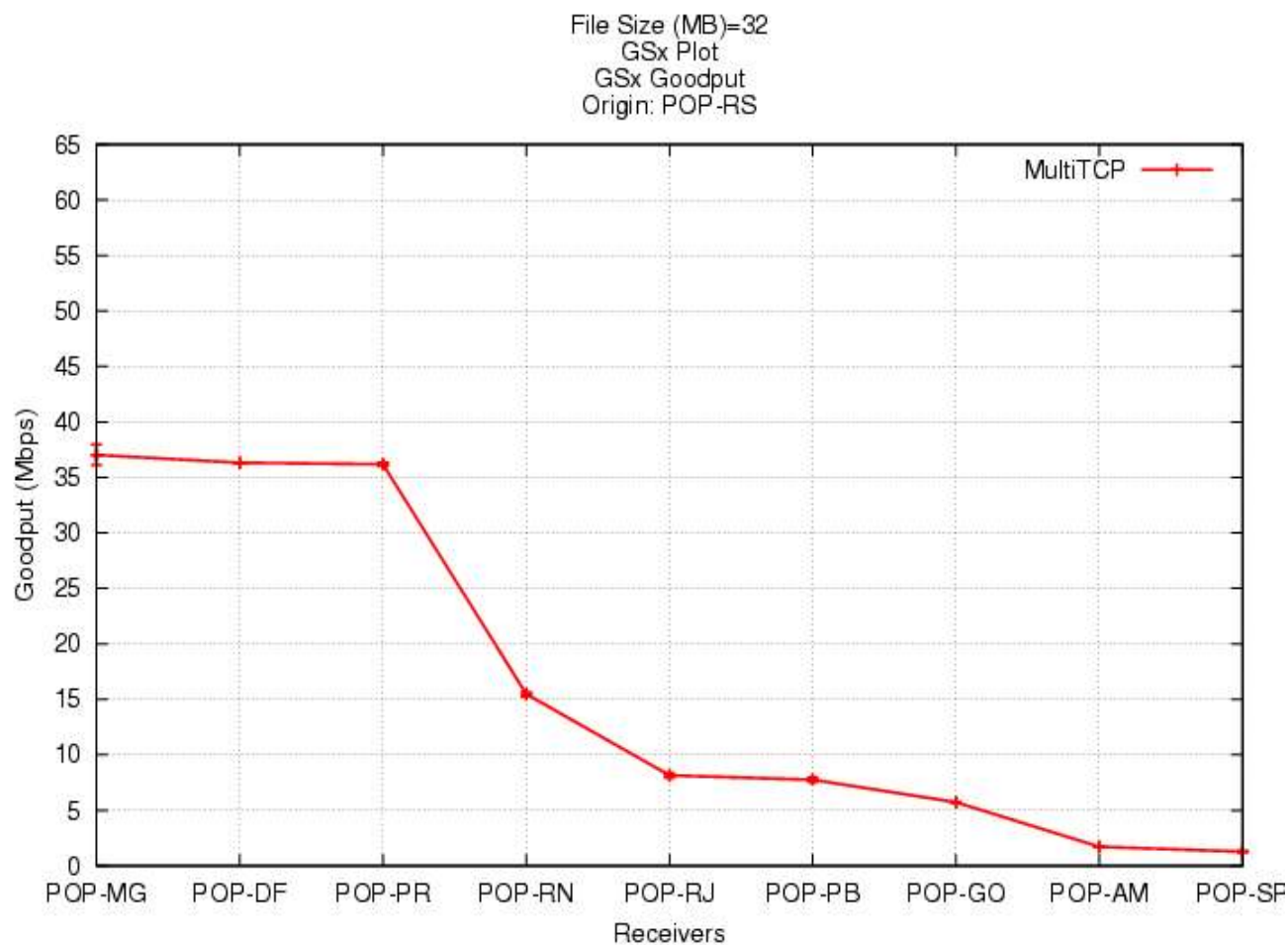
Preparação do ambiente 3

- Definição da ordem das máquinas em relação à taxa de transferência
 - Crucial para definir a ordem de crescimento dos grupos multicast
 - Caso não estivesse em ordem, o POP mais lento iria mascarar os resultados dos POPs mais rápidos



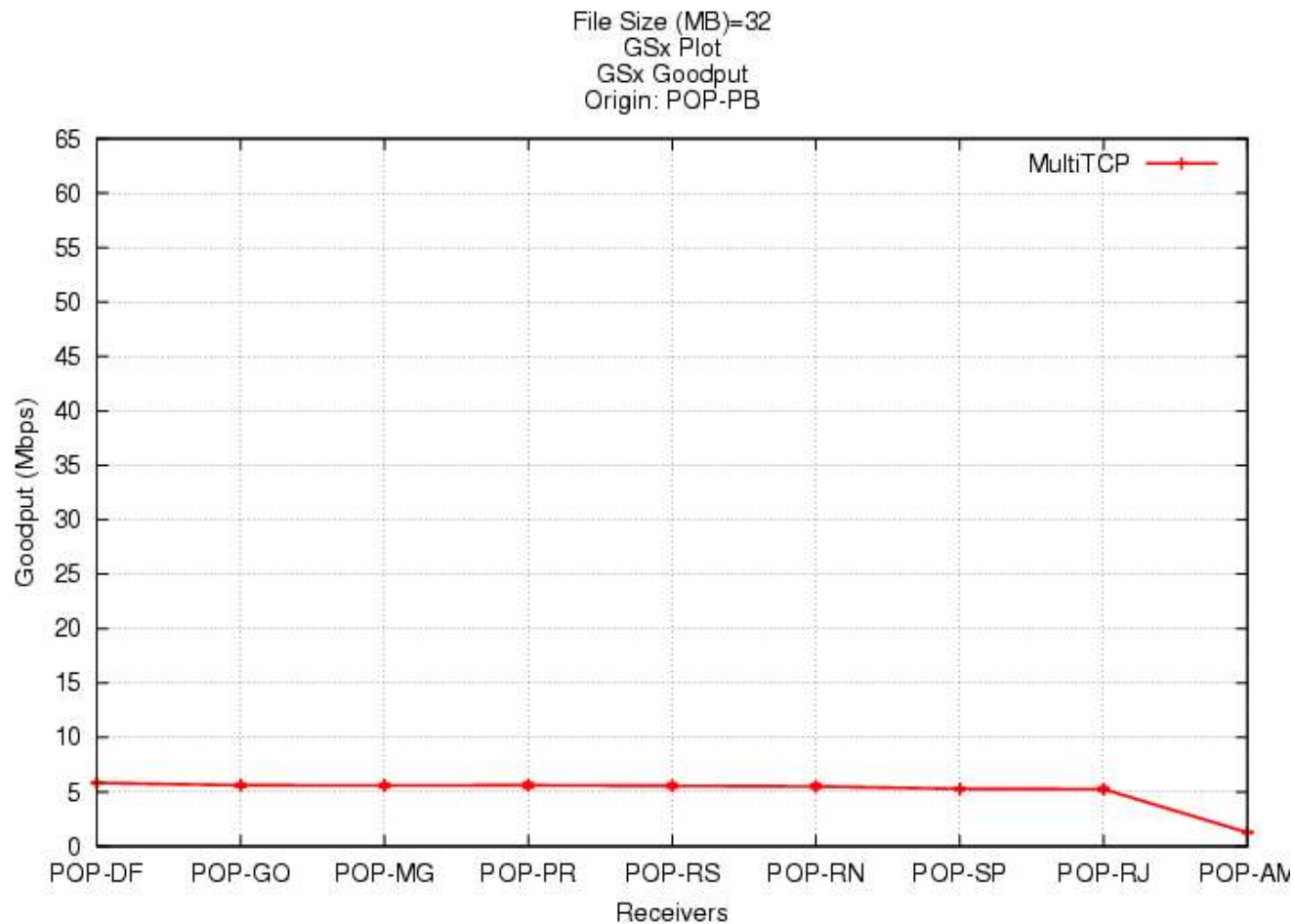
Taxas 1:1 – Origem RS

- RS-SP: Buffers de 4MBytes; RTT 70ms; link de 155Mbps/s com 30% utilização; horário entre 1 e 6 da madrugada
- Taxa baixa ???



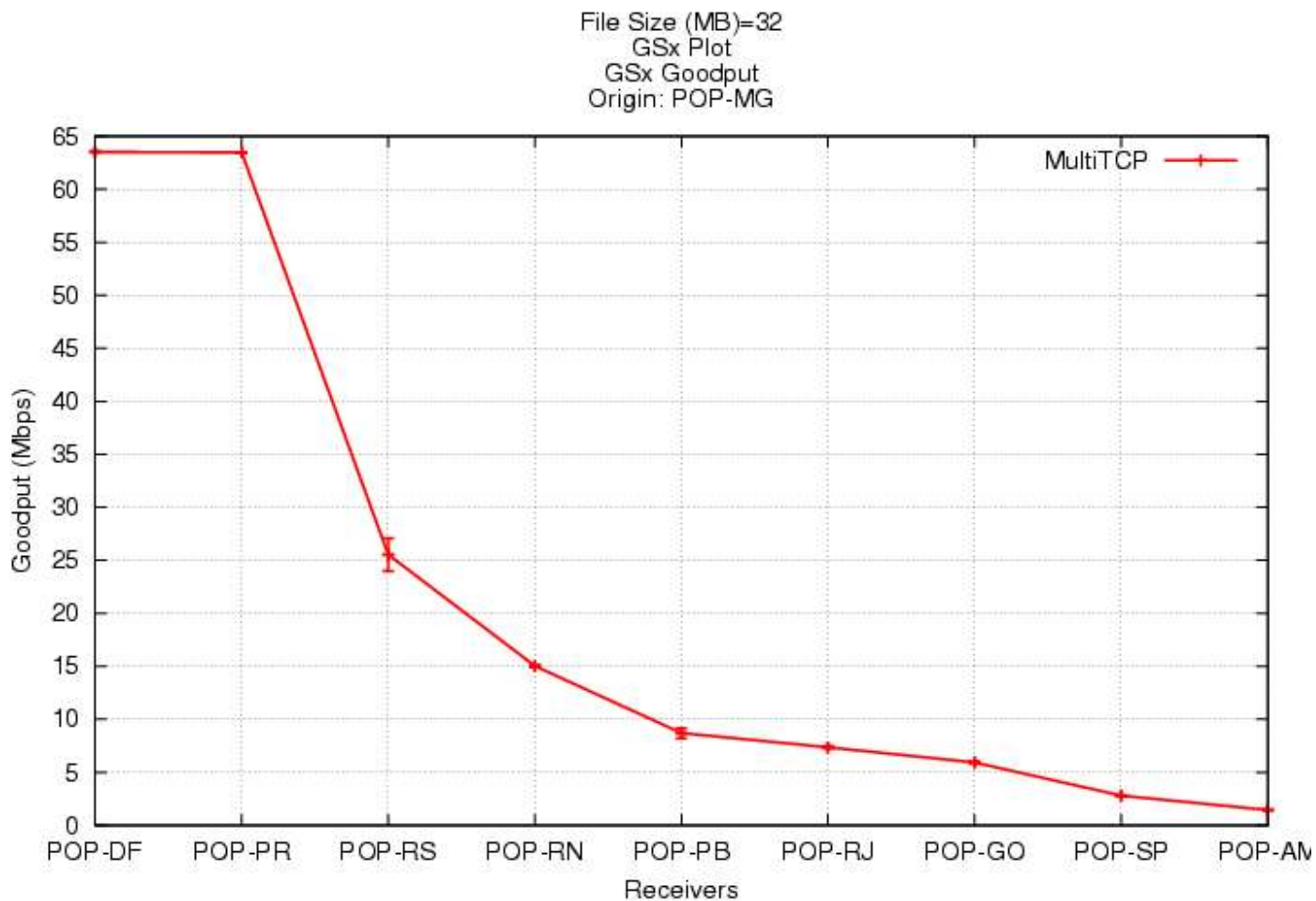
Taxas 1:1 – origem PB

- Percebe-se gargalo no link de saída do POP-PB
- PB-SP com 5Mbit/s; RS-SP menos de 2Mbit/s ???



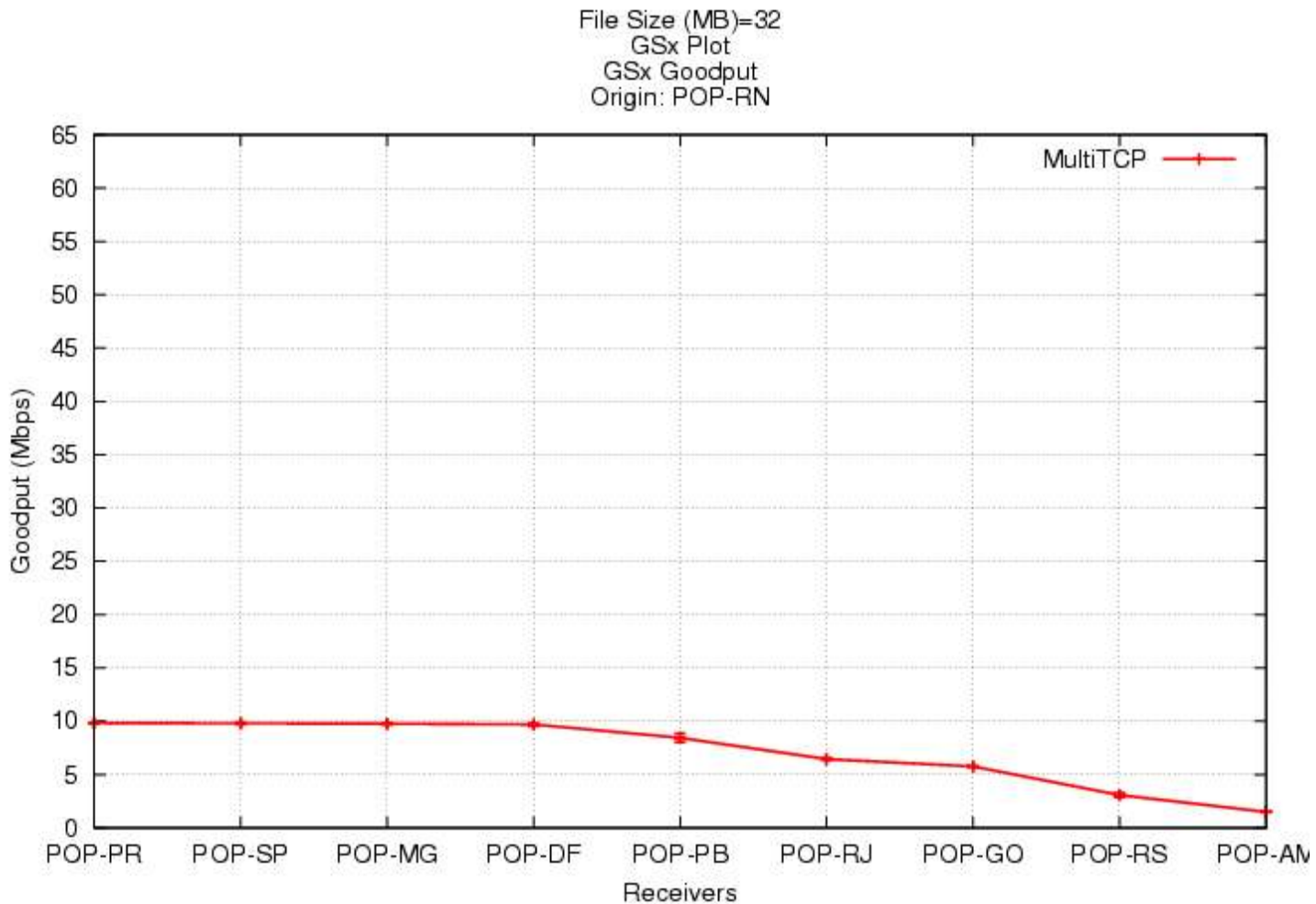
Taxas 1:1 – origem MG

- Ligado diretamente no POP-RJ, mesmo assim, desempenho baixo (7,5 Mbit/s)



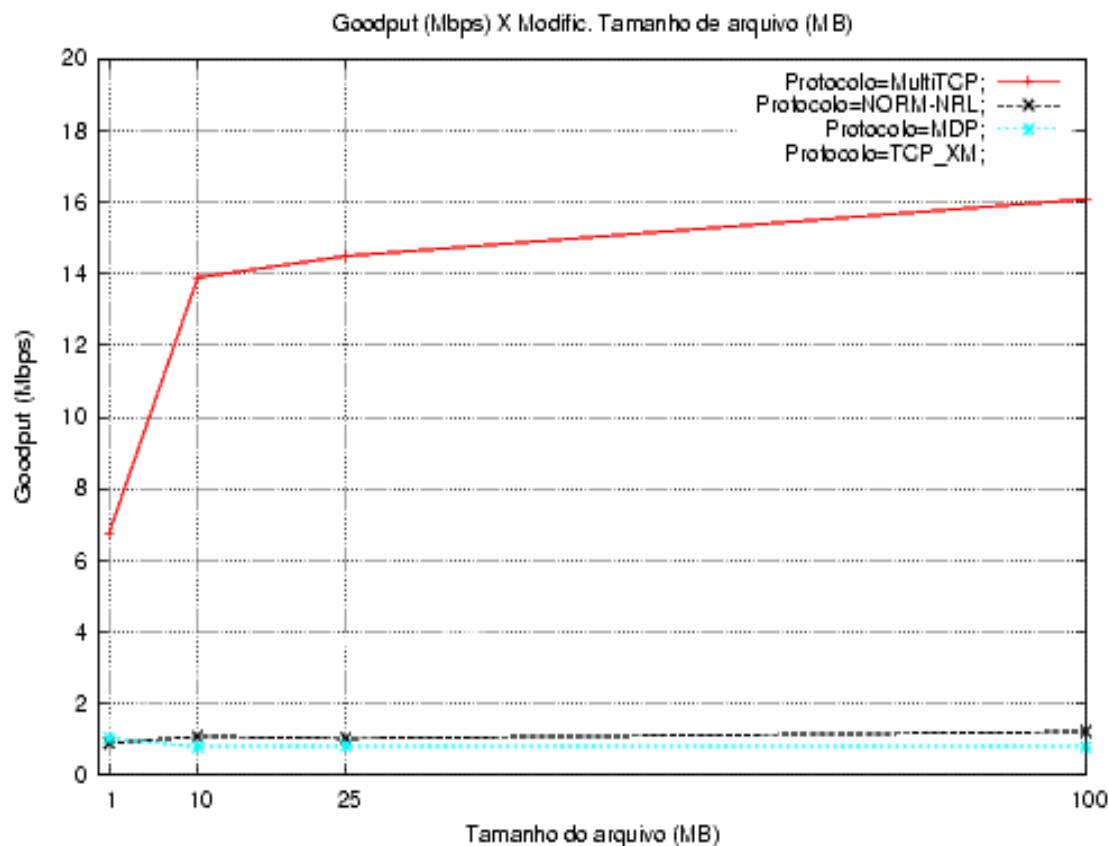
Taxas 1:1 – origem RN

- Gargalo no link do POP-RN
- POP-RS com desempenho baixo ???



Preparação do ambiente 3

- Definição do tamanho do arquivo
 - Variou-se o tamanho do arquivo para verificar o mínimo que poderia ser utilizado (diminui tempo dos experimentos)
 - Concluiu-se que os protocolos já estão estabilizados para arquivos com 10 Mbytes ou mais



Preparação do ambiente 4

- Definição dos parâmetros de FEC para protocolos
 - Utilizou-se um busca esparsa no espaço multidimensional que compõem todas as possibilidades de configuração de FEC. Foram feitos cerca de 500 experimentos que partiram do POP-RS (o único ponto de partida estudado devido ao alto tempo consumido para realizar esse tipo de experimento). Os melhores resultados foram colocados em uma tabela
- Detectou-se que o nível de FEC utilizado impacta bastante no uso do processador da máquina e no desempenho da transmissão
 - O ideal seria que o protocolo reajustasse o nível de FEC dinamicamente, pois condições de rede podem mudar durante a transmissão. Como isso não acontece, devemos tentar 'prever' os melhores valores para uma situação prática e utilizá-los.
- Concluiu-se que os melhores parâmetros de FEC seriam
 - Tamanho do bloco: 95
 - Paridade extra calculada: 32
 - Envio pró-ativo: 16
 - Extra (em caso de perdas): 4



Descrição dos Experimentos

- Envio de um arquivo para um conjunto de máquinas destino usando um protocolo de transmissão multicast confiável
- Coleta de resultados de custo (largura de banda) e desempenho (goodput)
- Experimentos prévios:
 - avaliação do impacto do tamanho do arquivo
 - determinação de configuração dos protocolos
 - determinação da ordem ideal das máquinas



Ambientes de Redes Avaliados

- Rede Local Fast Ethernet
- Agregado Gigabit Ethernet
- RNP
- SuperJanet



Laboratório de rede local utilizado

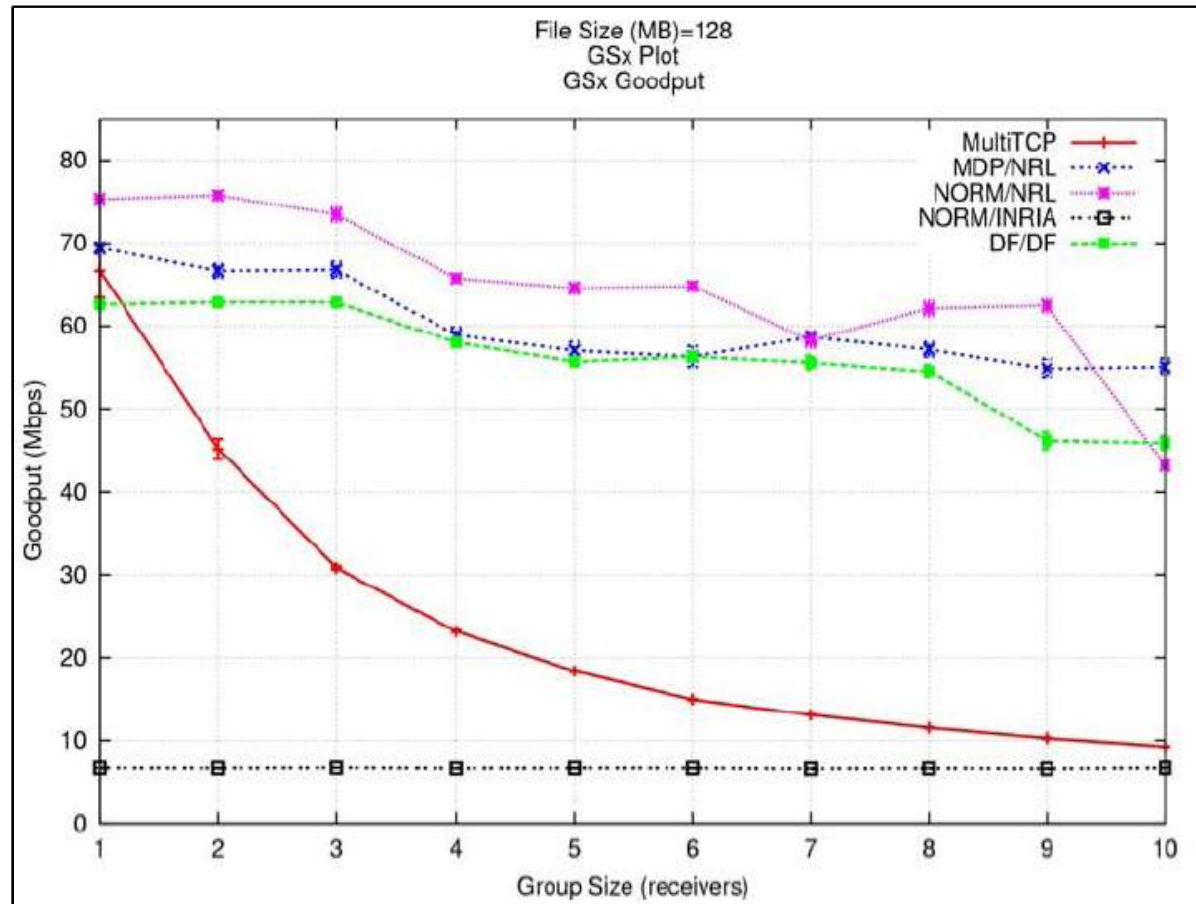
- Utilizado grupo variável de 1 até 10 receptores
- 4xPIV 2,4 Ghz (1GRam) e 7x1,8 Ghz (256MRam)



Resultados na Rede local

Taxa fixa de 100 Mbps

- Topologia com Switch (transmissor a 100Mbps)
- Overhead do TCP ainda aumenta com o n. clientes



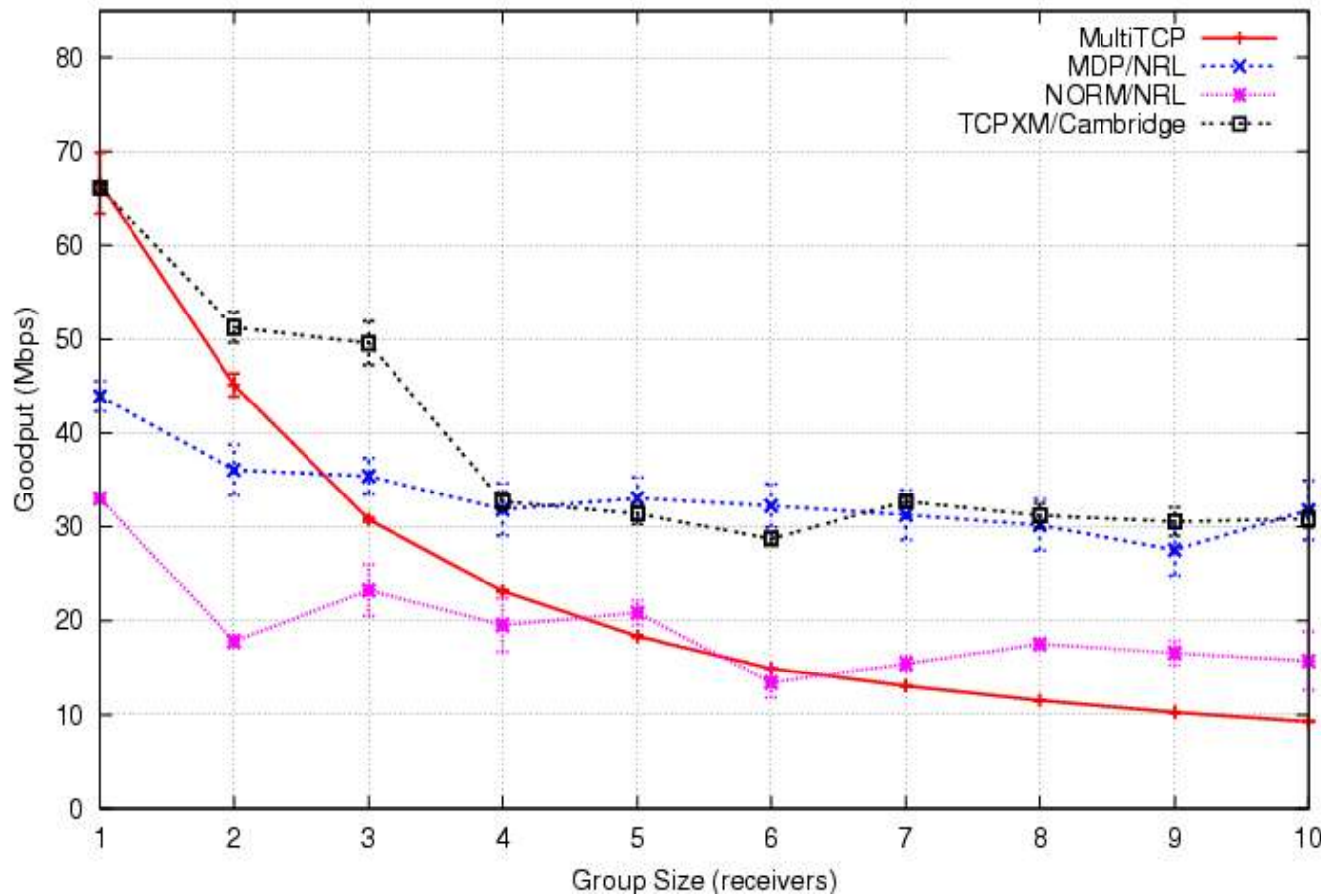
OBS: falta inserir gráfico de overhead TCP junto

- NORM/INRIA: tem limite de 10Mbit/s – não é adequado para redes de alta velocidade

Resultados na Rede Local

Taxa Adaptativa

File Size (MB)=128
Local Network - Adaptative
Group Size X Goodput



OBS: falta inserir gráfico de overhead TCP junto

Agregado giga utilizado

- Utilizado grupo variável de 1 até 10 receptores
- 5x Xeon Dual 2,4GHz (2GRam), 6x Xeon Dual 2,8 Ghz (2G Ram)

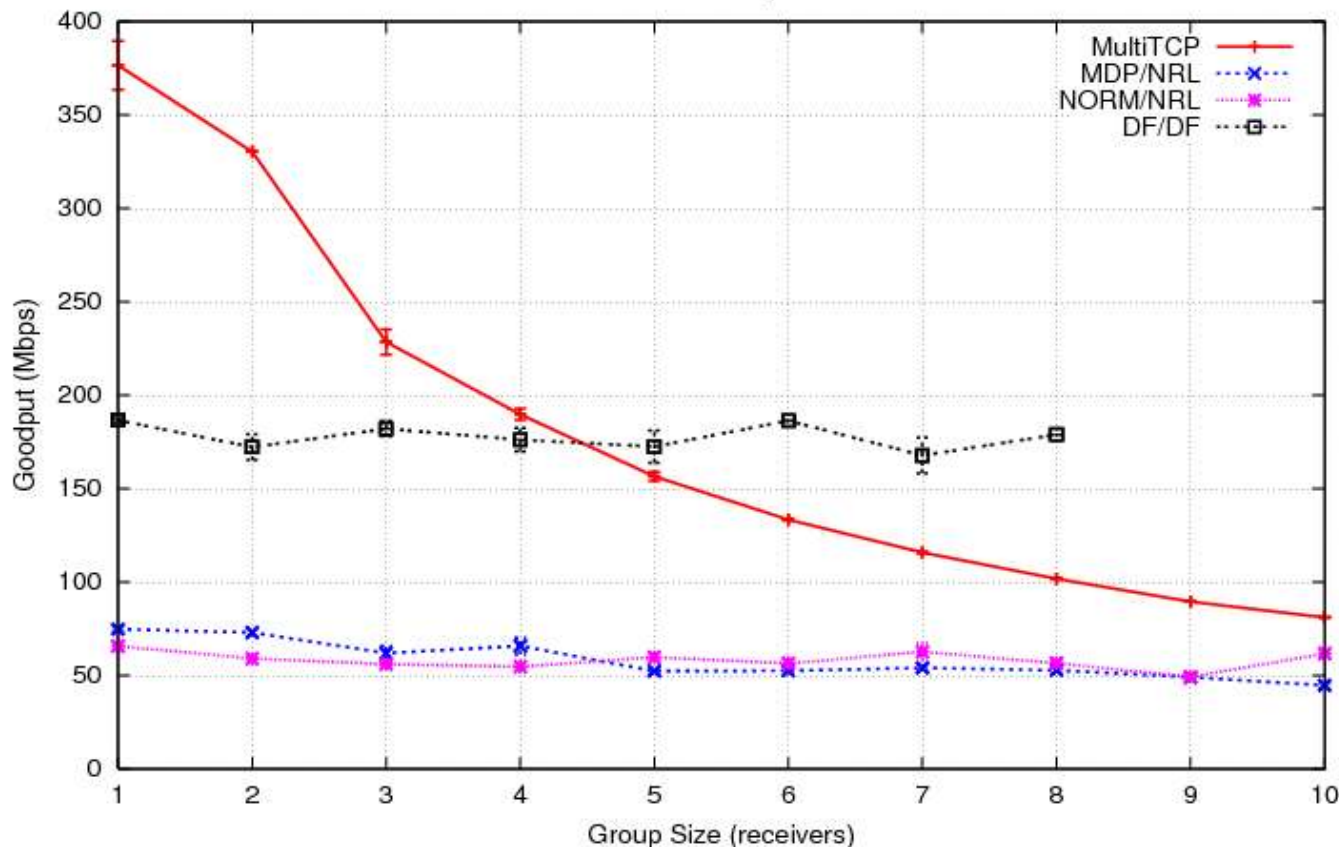


Resultado no Agregado Giga

Taxa Fixa de 800 Mbps

- Protocolos não passam de 100 Mbit/s (menos o DF)
- Multi-tcp: overhead aumenta com número de clientes

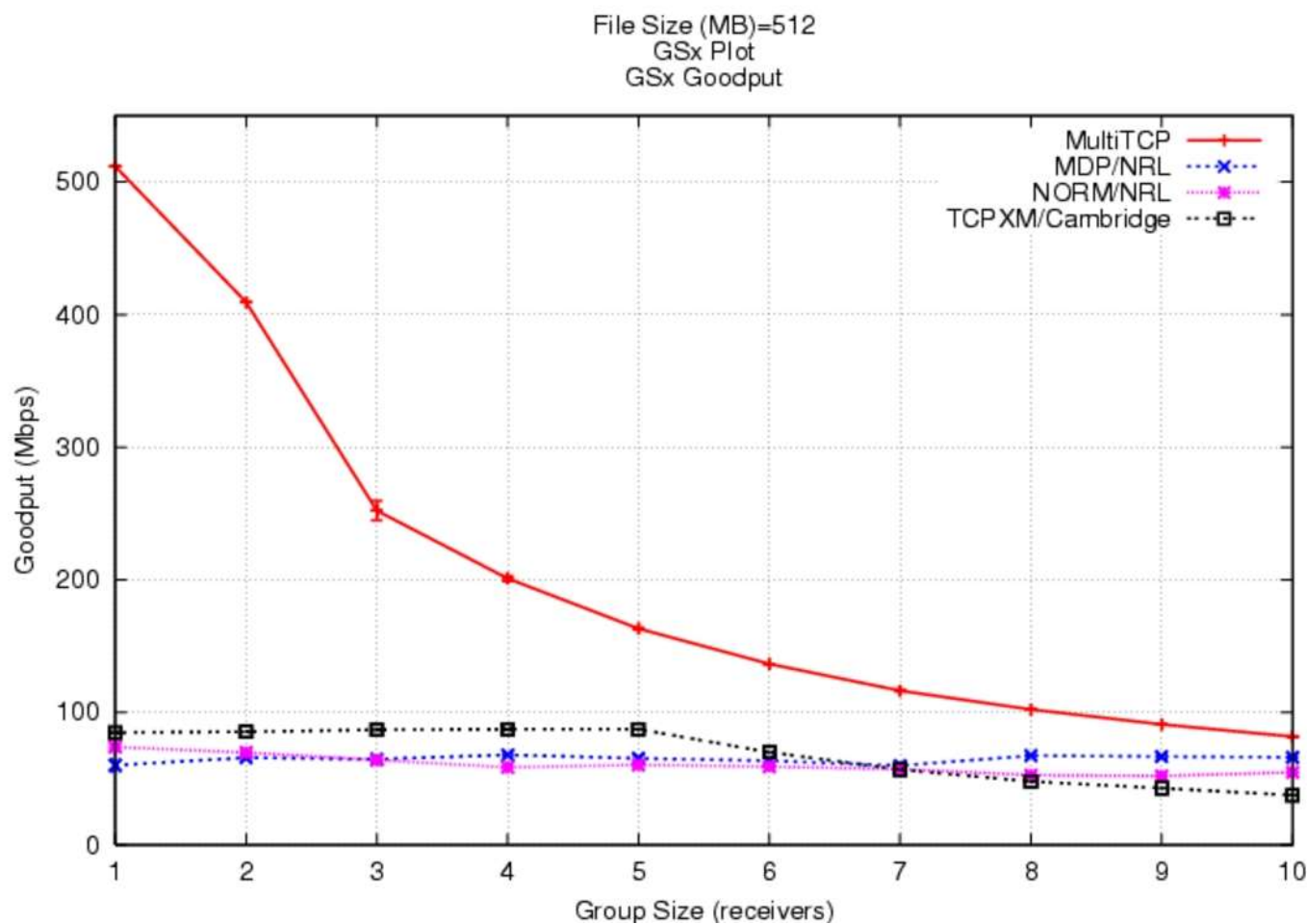
File Size (MB)=536
GSx Plot
GSx Goodput



OBS: falta inserir gráfico de overhead TCP junto

Aggregado Giga Taxa Adaptativa

- Mais desempenho em relação à rede 100M
- Não passam de 100Mbit/s

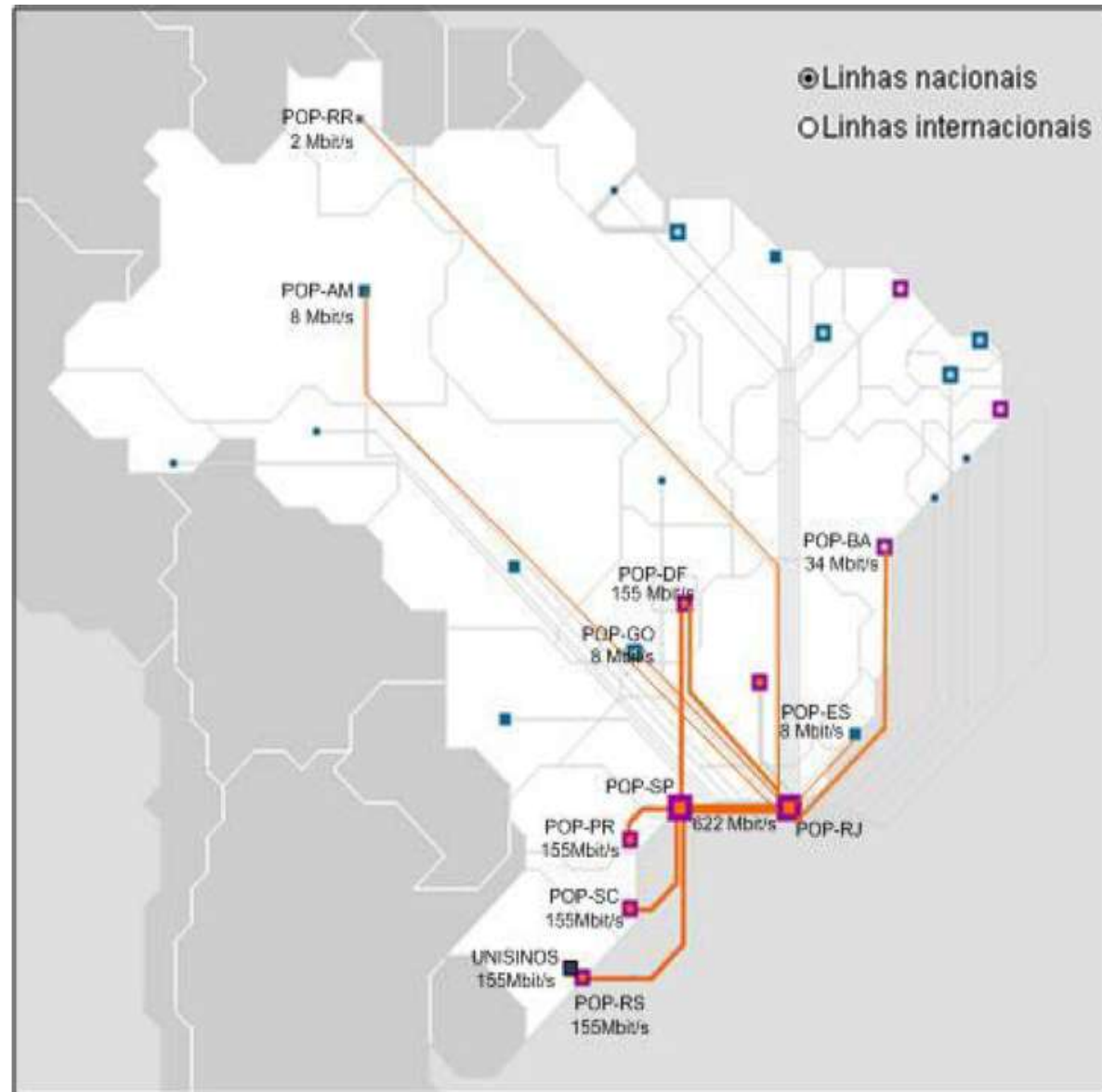


OBS: falta inserir gráfico de overhead TCP junto



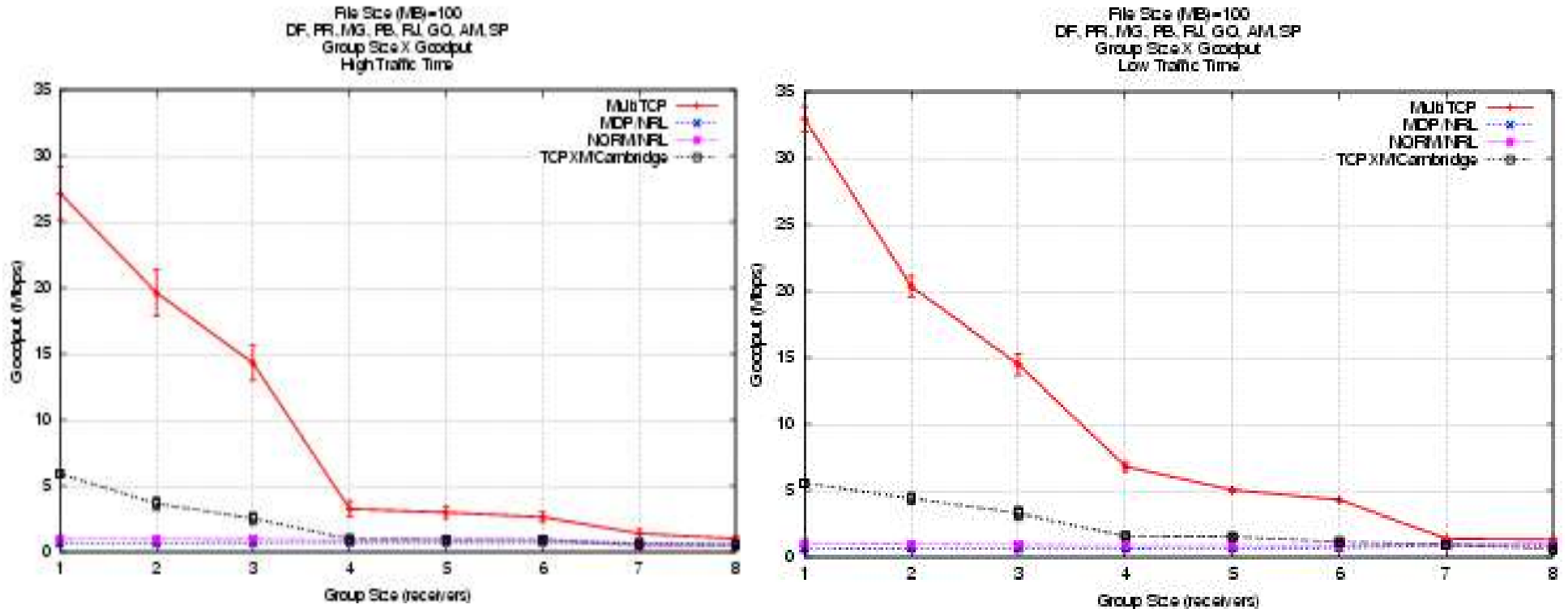
RNP

- POPs adotados:
 - RS, PR, SP, RJ, DF, MG, GO, PB, RN, AM
- POPs com problemas
 - SC e ES (sem multicast)
 - BA (sem espaço em disco)
 - RR (link lento e sobrecarregado)



RNP: transmissor RS

Taxa Adaptativa (goodput)



Resultados para rede no horário 'de pico' e madrugada. Interessante notar que não houve muitas mudanças, o que nos leva a crer que a rede está ociosa em muitos dos pontos.

- OBS: ordem das máquinas deve ser feita com UDP, pois com TCP não dá o mesmo resultado

RNP: transmissor RS

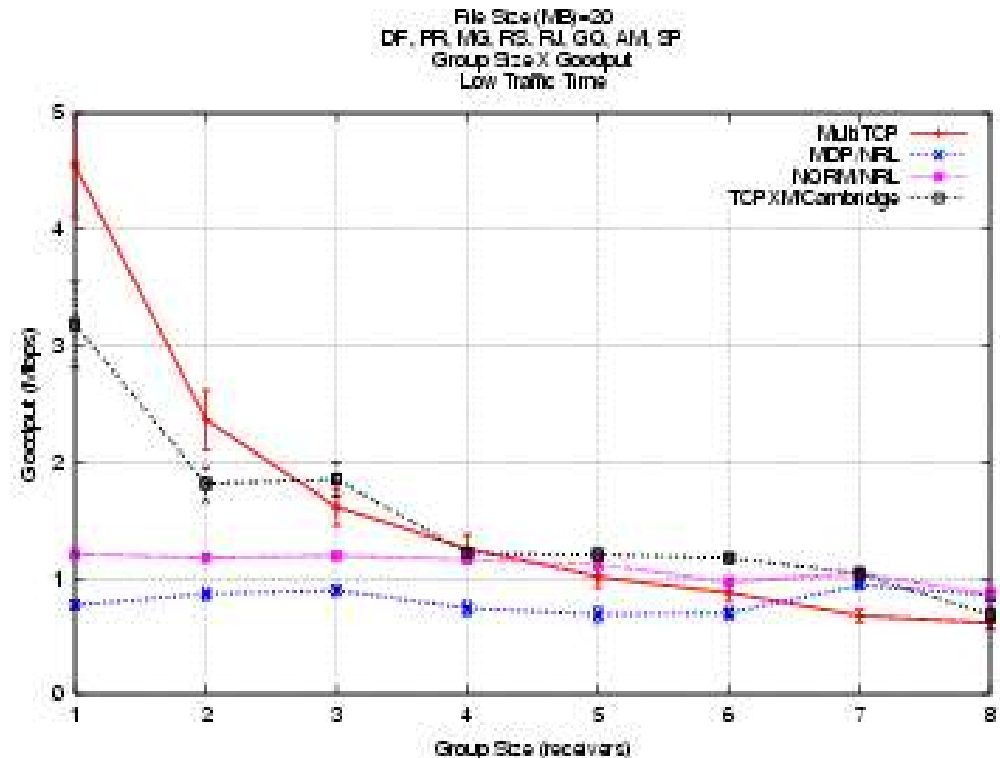
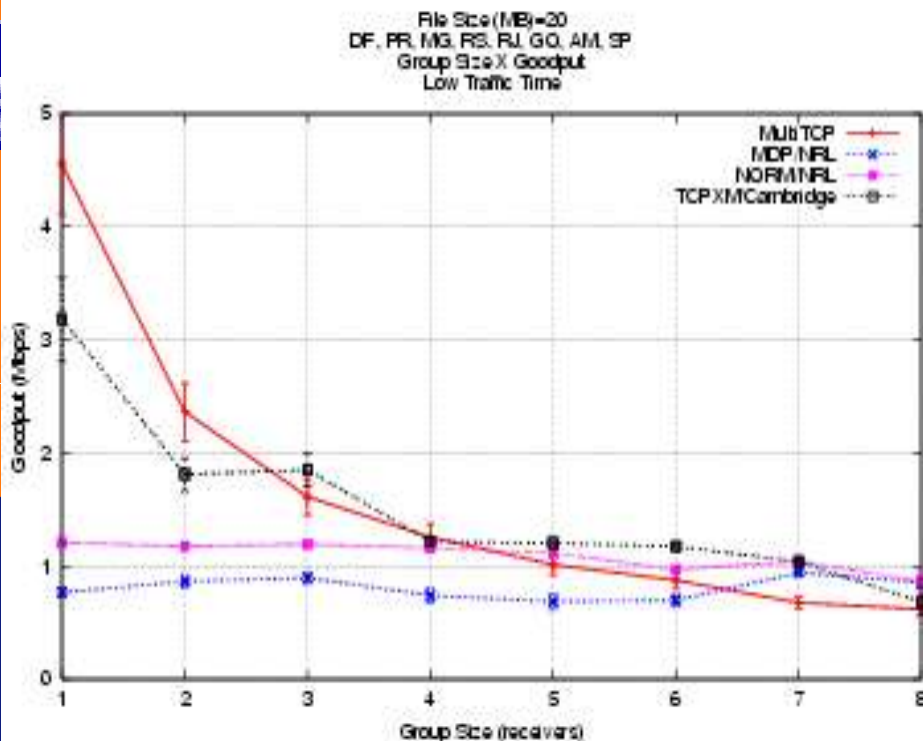
Taxa Adaptativa (overhead)

Falta gráfico de overhead TCP
Falta gráfico perdas no link



RNP: transmissor PB

Taxa Adaptativa (goodput)



- Pequena vantagem sobre o TCP (em termos de Goodput).
- Evidente característica conservadora do controle de congestionamento
- A vantagem é o menor overhead dos protocolos de multicast.

RNP: transmissor PB

Taxa Adaptativa (overhead)

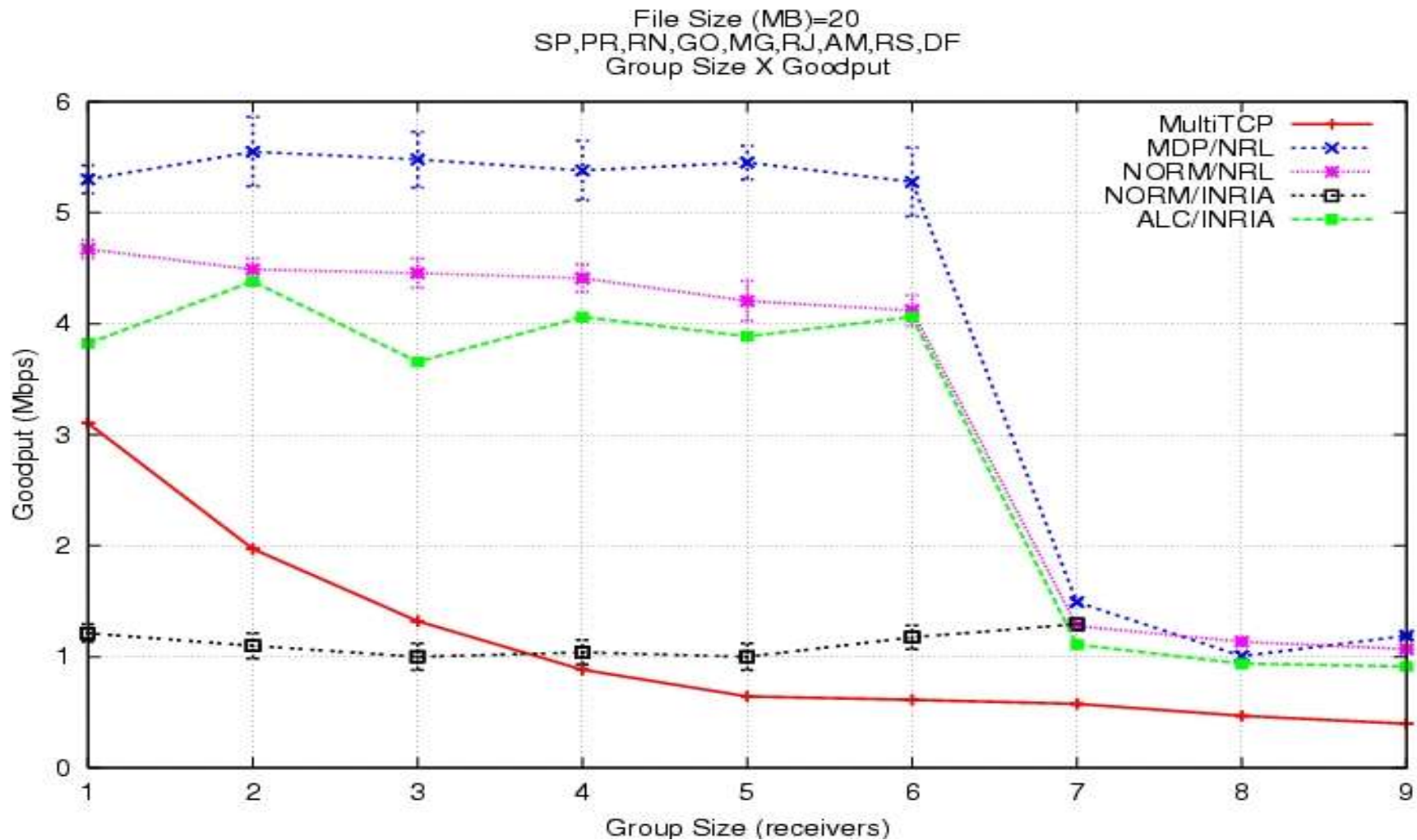
Falta gráfico de overhead TCP
Falta gráfico perdas no link



RNP: Transmissor PB

Taxa Fixa (goodput)

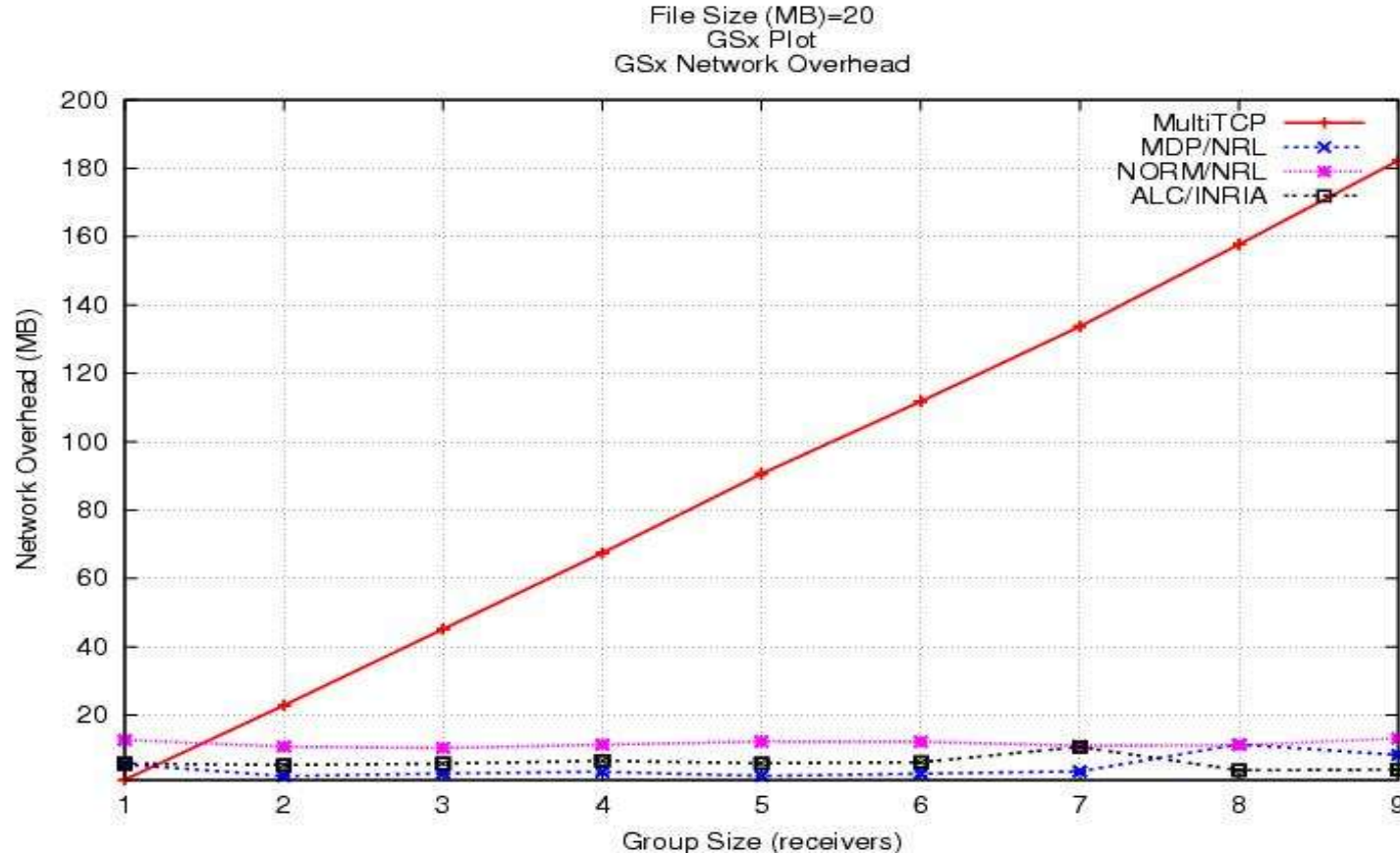
- Taxa fixa baseada no menor goodput multicast do grupo



RNP: Transmissor PB










Taxa Fixa (sobrecarga gerada)

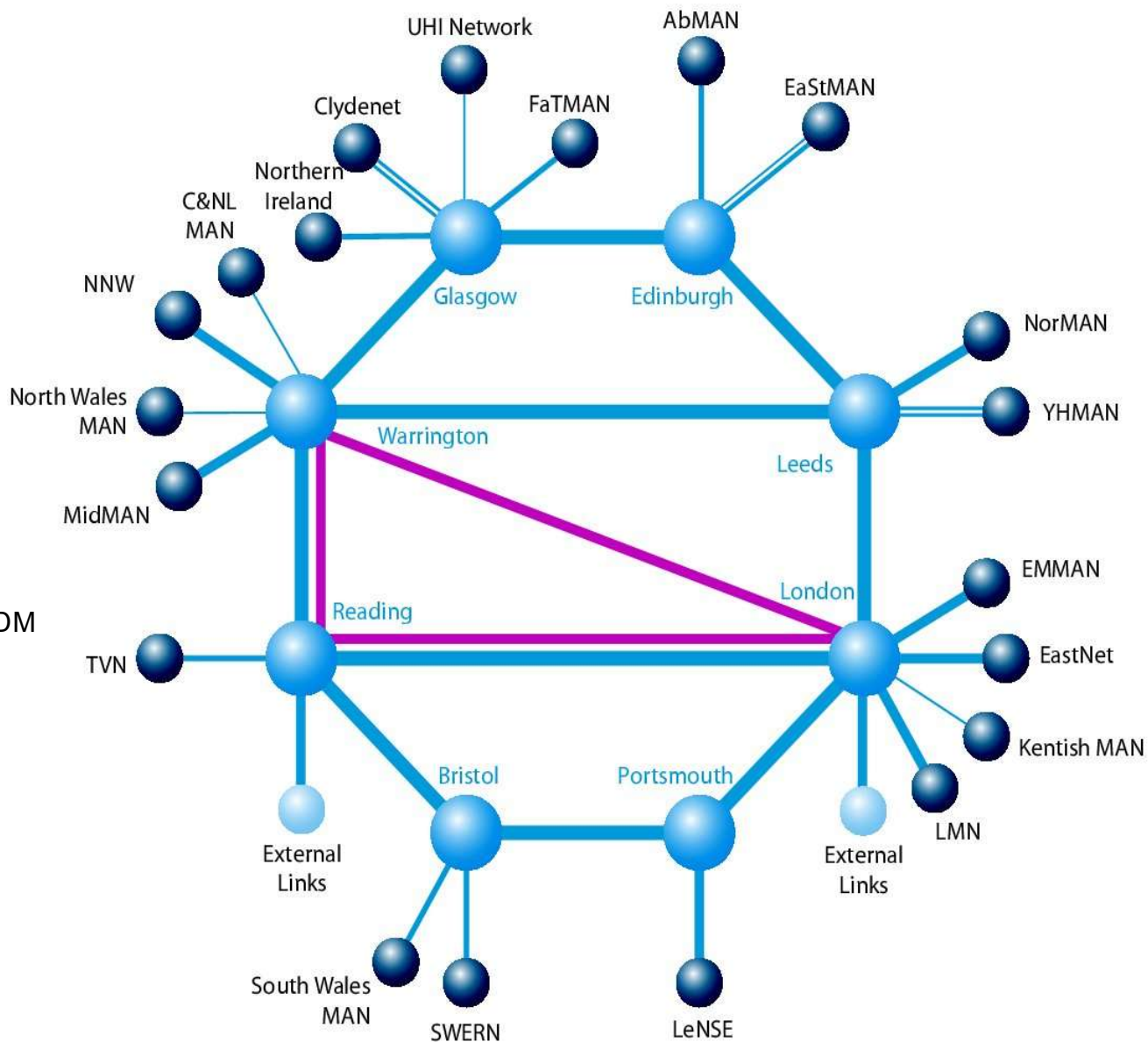
- O gráfico mostra o problema do multi-unicast, que é justamente a sobrecarga gerada na rede



SuperJanet



-  Core POP Router
-  Backbone Access Router
-  External Links
-  Backbone link - 10 Gbit/s DWDM
-  Test-bed Network - 2,5 Gbit/s DWDM
-  Dark fibre at 2,5 Gbit/s
-  2,5 Gbit/s SDH
-  622 Mbit/s SDH
-  155 Mbit/s SDH

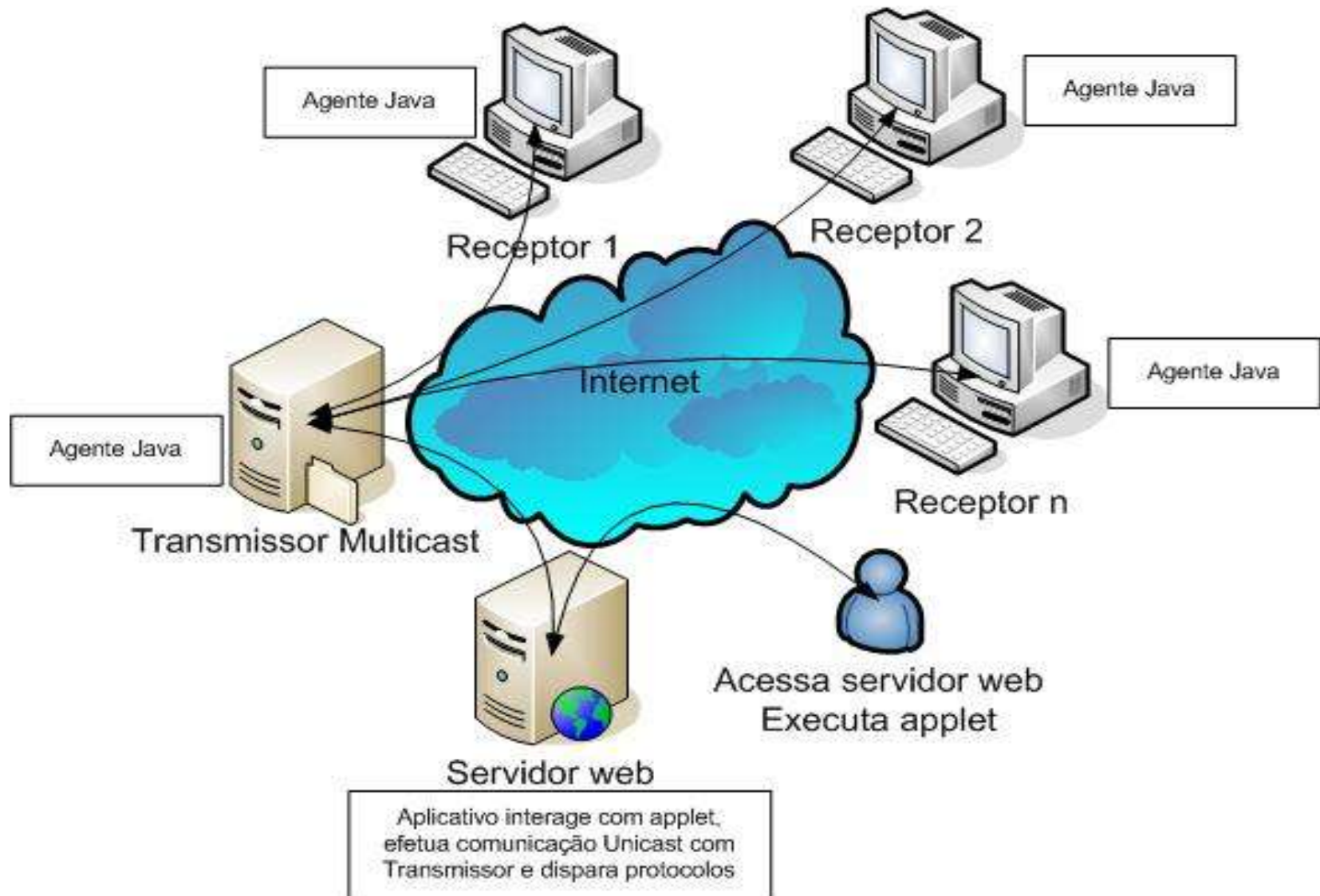


Experimentos: Conclusões Parciais

- Foco não é o desempenho, e sim a eficiência (diminuição do volume de tráfego com multicast)
- Abandonados
 - JRMS: implementação instável e sem suporte
 - NORM/INRIA: implementação instável e sem controle de congestionamento
 - DF: protocolo comercial e proprietário. Caso taxa de perdas maior que FEC utilizado, transmissão falha
- Candidatos são
 - MDP: desempenho
 - NORM/NRL: controle de congestionamento amigável ao TCP
 - ALC/MCL: heterogeneidade de taxas via camadas
 - TCP-XM: híbrido uni-multicast, amigabilidade ao TCP



Interface com o Usuário



Interface com o Usuário

Agendamento de transferência | Gerenciamento

Transmissor: Arquivo de origem: ...

Diretório de destino:

Receptores potenciais

- pop-df
- pop-es
- pop-pr
- pop-sp
- pop-rs
- pop-go
- pop-rr
- pop-am

Perfil de protocolo

Horário de transferência

Dia: / /

Horário: :

Enviar agora

Principais Problemas Encontrados

- Perda de home no POP BA;
- Instabilidade em algumas interconexões: principalmente nos uplinks dos POPs da região Norte, prejudicando determinados experimentos que tiveram que ser reiniciados. Entre eles ES e DF.
- Espaço em disco: Alguns POPs (BA e DF) tiveram todo espaço em disco ocupado, cancelando diversos experimentos.
- O protocolo TCP-XM não está funcional nos POPs RR e RN. Apesar de várias tentativas não foi descoberta a causa do problema.
- Não existe multicast habilitado para com os POPs SC e ES.
- **OBS: suporte ágil da equipe RNP**



Aplicações analisadas

- Aplicação 1: Transferência de arquivos de vídeo entre TVs Universitárias (foco principal)
- Aplicação 2: Jogos MMG (*Massively Multiplayer Games*)



Aplicação 2: jogos distribuídos em rede

- Características

- MMG: *Massively Multiplayer Games*
- Grande quantidade de participantes simultâneos
- Envio freqüente de mensagens pequenas via rede

- Não é adequado para multicast confiável

- Multicast confiável não se aplica muito bem, pois, para arquivos pequenos, o protocolo praticamente não chega a estabilizar
- Jogos tem muitas necessidades como: baixa latência (para interatividade), baixo jitter (para não se perder a noção de quanto tempo uma ação do jogador vai realmente ocorrer).
- Multicast confiável não tem o desempenho que um jogo necessita (baixa latência)



Conclusões e Próximos Passos

- Execução e análise dos resultados de forma mais extensiva
- Refazer ordem das máquinas com UDP e experimentos com a nova ordem
- Geração do relatório final
- Transferência de tecnologia para a RNP
- Publicações
- Aprimoramento da interface com o usuário (estatísticas)
- Extensão do estudo sobre multicast confiável para a nova rede da RNP
- Explorar novas aplicações de multicast, como transmissão em tempo real (com e sem interatividade)



Implantação e Avaliação de Desempenho de Protocolos de Transmissão Multicast Confiável na RNP

PERGUNTAS?

**Valter Roesler, Marinho P. Barcellos
Evandro C. Dall’Agnol, Giovani Facchini
Gustavo Bervian Brand, Renato Costa,
Tasso Gomes de Farias**

Universidade do Vale do Rio dos Sinos (UNISINOS)
Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
PRAV – Laboratório de Redes de Alta Velocidade

<http://prav.unisinos.br/gtmc>

