

ネットワークトランスポートの  
研究開発 -Grid 環境の実現に向  
けて-

独立行政法人 通信総合研究所

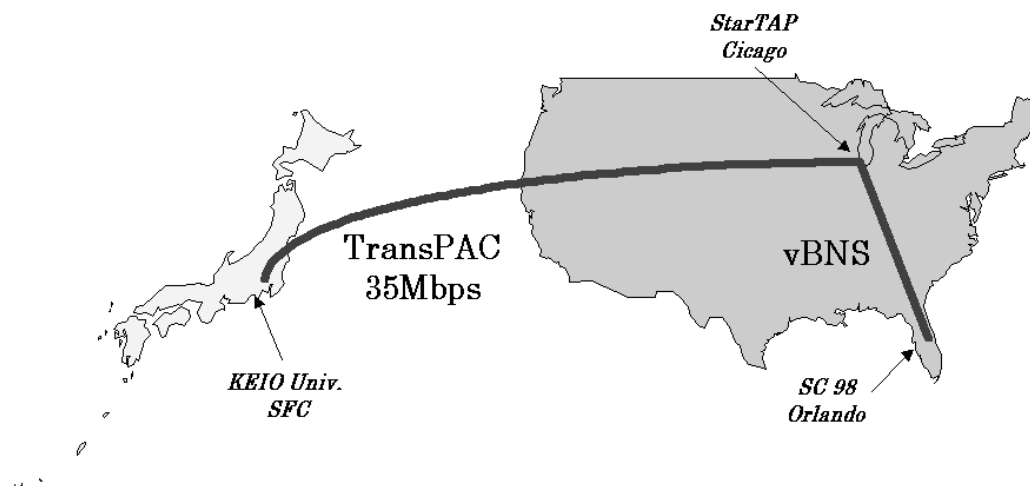
小林 克志

# Agenda

- はじめに
- マルチキャストの展開
- 
- TCP 性能計測プラットフォームの展開
- トランスポートの研究開発

# はじめに

- Grid との関わり
  - SC'98 (Orlando) において TransPAC ネットワークのお披露目デモを iGrid 展示として行った。
  - DV over IP を遠隔電子顕微鏡実験に提供



<http://www.sfc.wide.ad.jp/DVTS/sc98.html>

# マルチキャストの展開

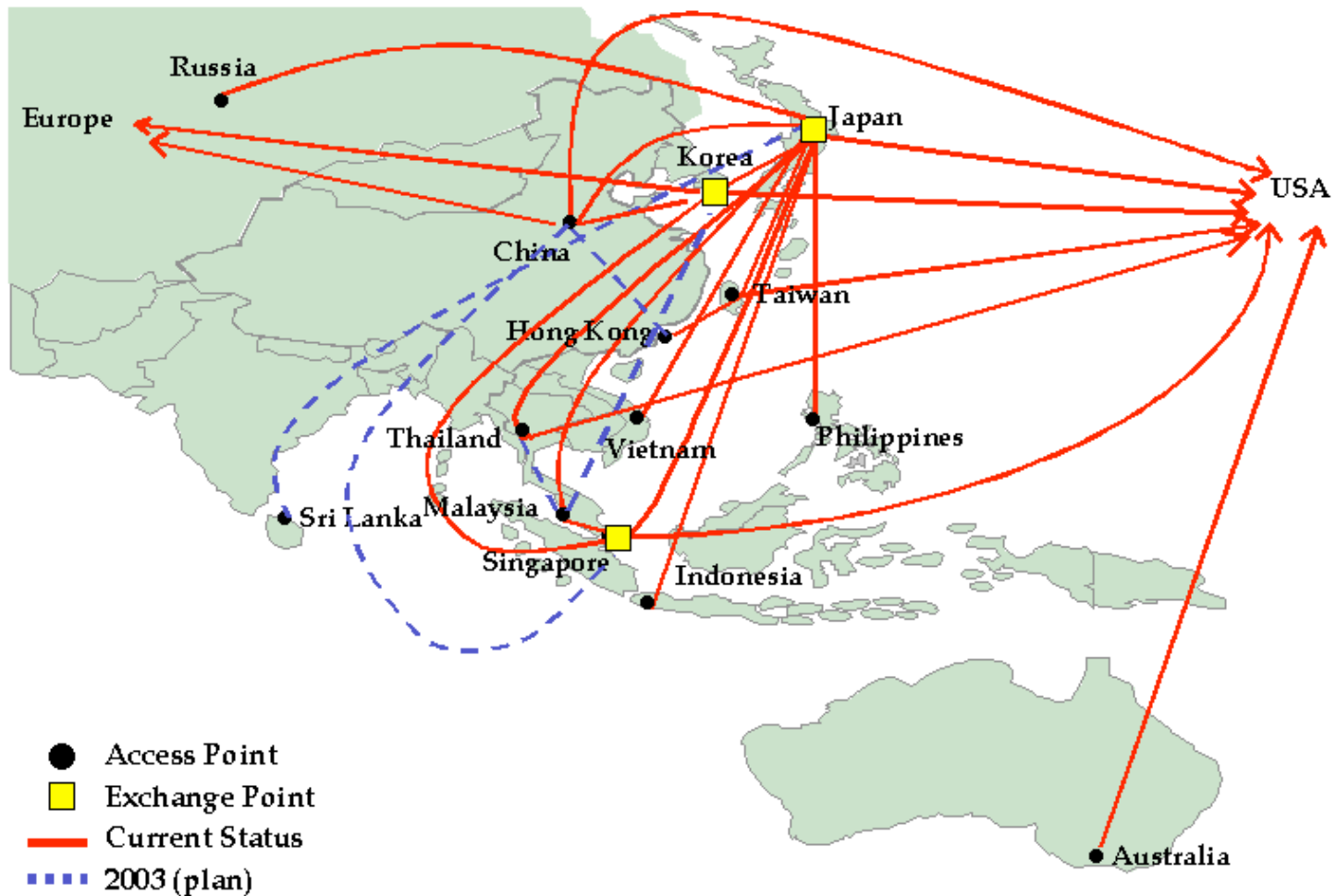
- アクセスグリッド展開にはマルチキャスト基盤の展開が不可欠
- US, 欧州の研究開発バックボーンでは、Native マルチキャストへの移行は終了しているが、日本では依然として Mbone ベースのネットワーク
- 世界的なマルチキャストネットワークの展開規模の評価

# マルチキャストの展開

- TransPAC の国内接続点、APAN TokyoXP で、マルチキャスト情報の収集、解析
- Why TokyoXP
  - 国内の大半のマルチキャスト網の対外接続点
  - APAN 傘下のネットワークを介して、CN, KR, TW へも接続提供、アジア地域の接続点として重要
- 期間 2003 年 2 月

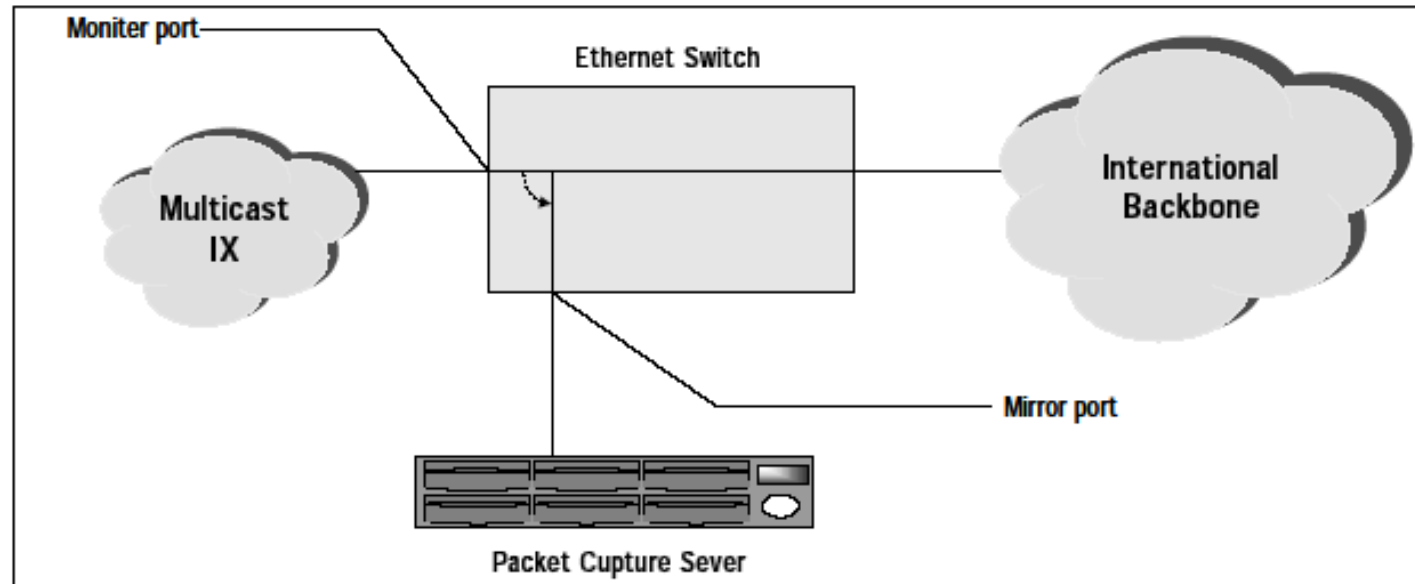
# マルチキャストの展開

APAN Network Topology (updated 2003. 1. 15)



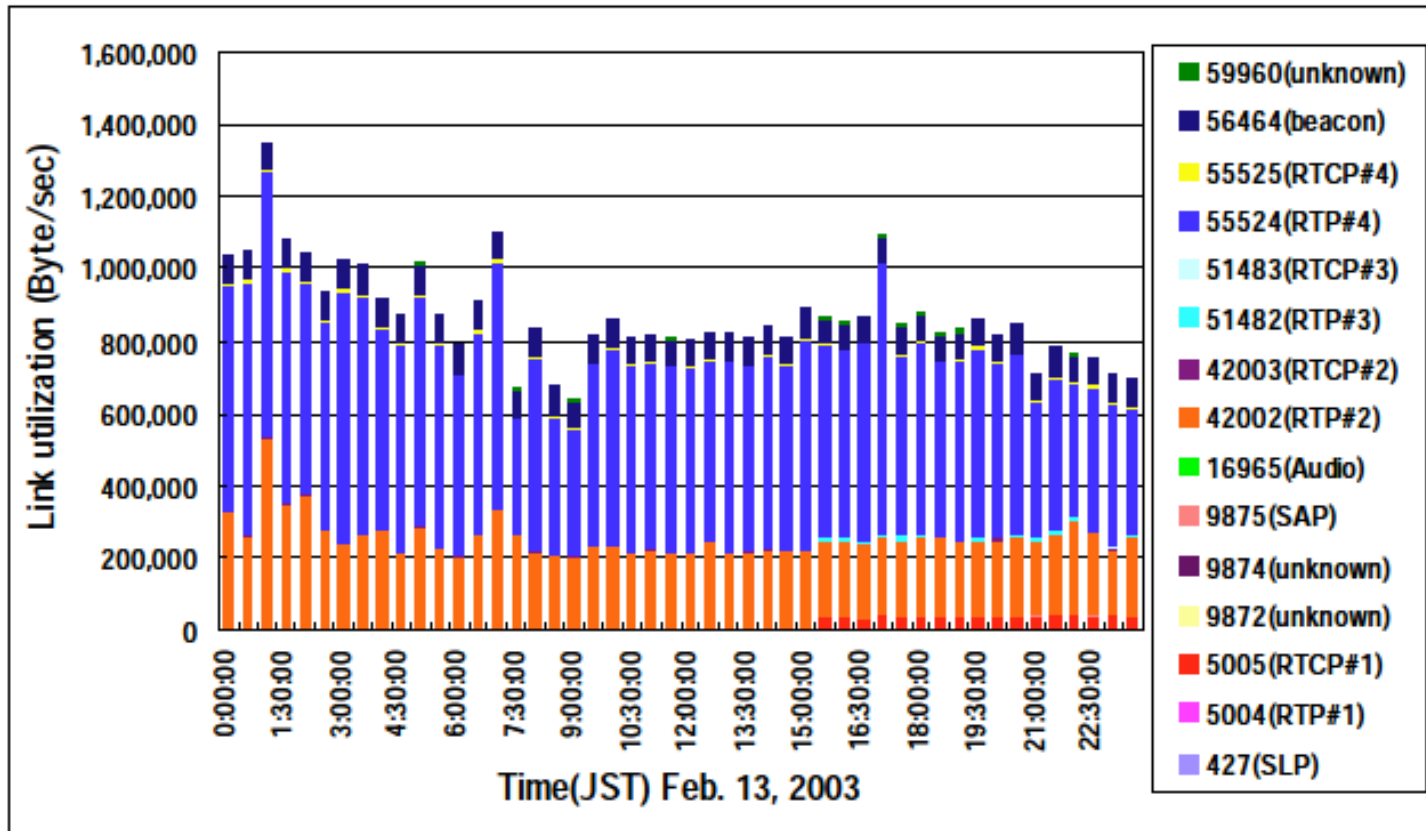
# マルチキャストの展開

- APAN TokyoXP で国際接続 - 国内接続  
の中間でデータ取得、分析



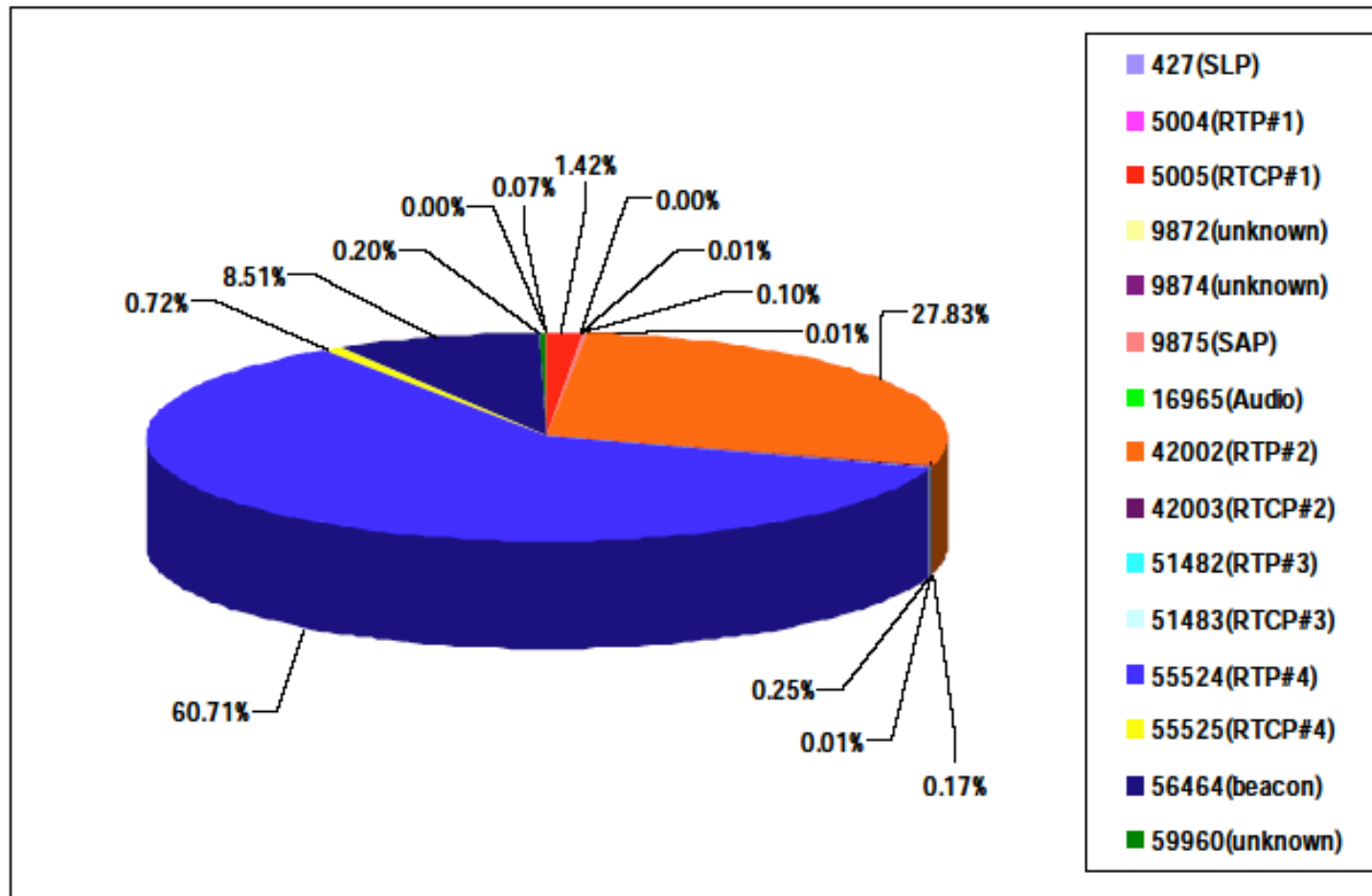
# マルチキャストの展開

## 1.4.3 マルチキャストアプリケーション毎の利用帯域

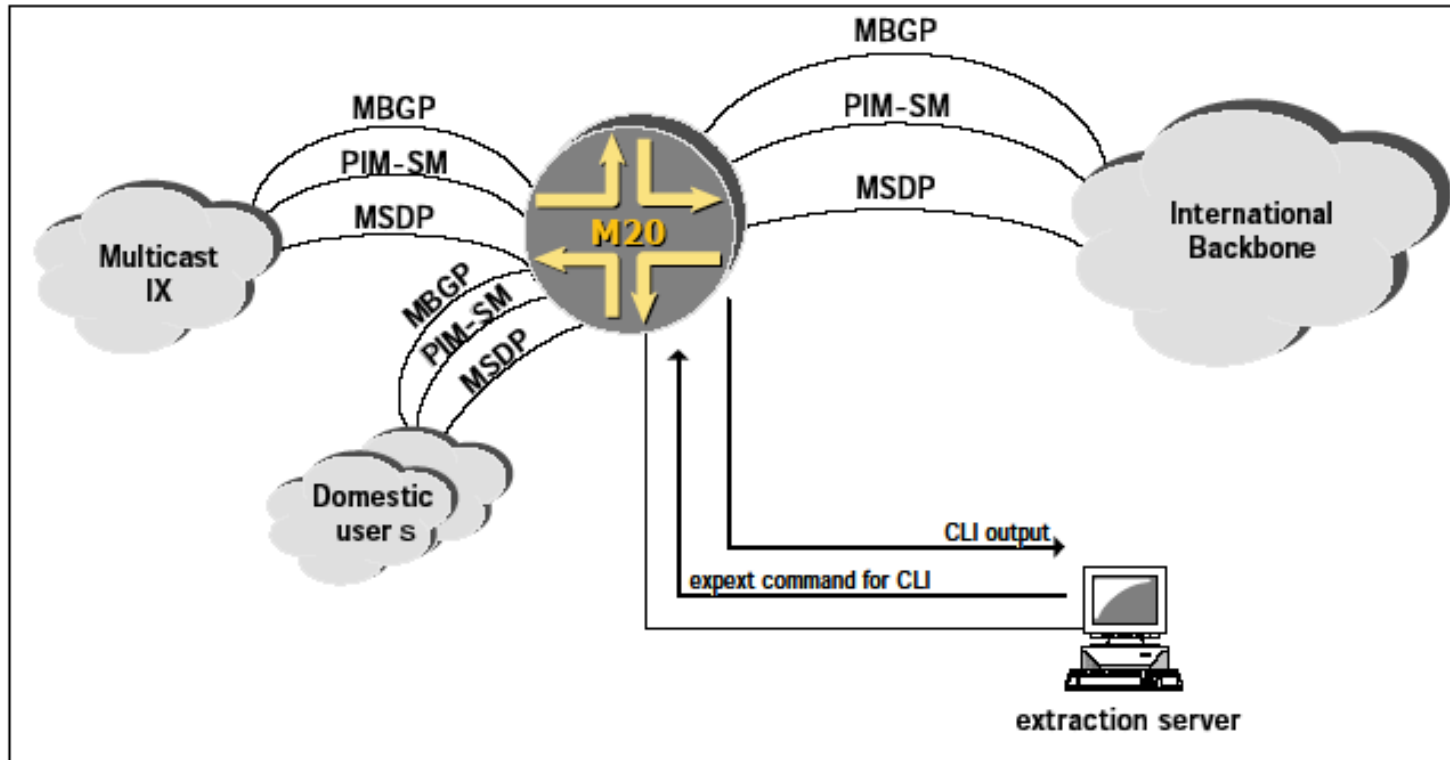


# マルチキャストの展開

## 1.4.4 マルチキャストアプリケーションの割合



# マルチキャストの展開

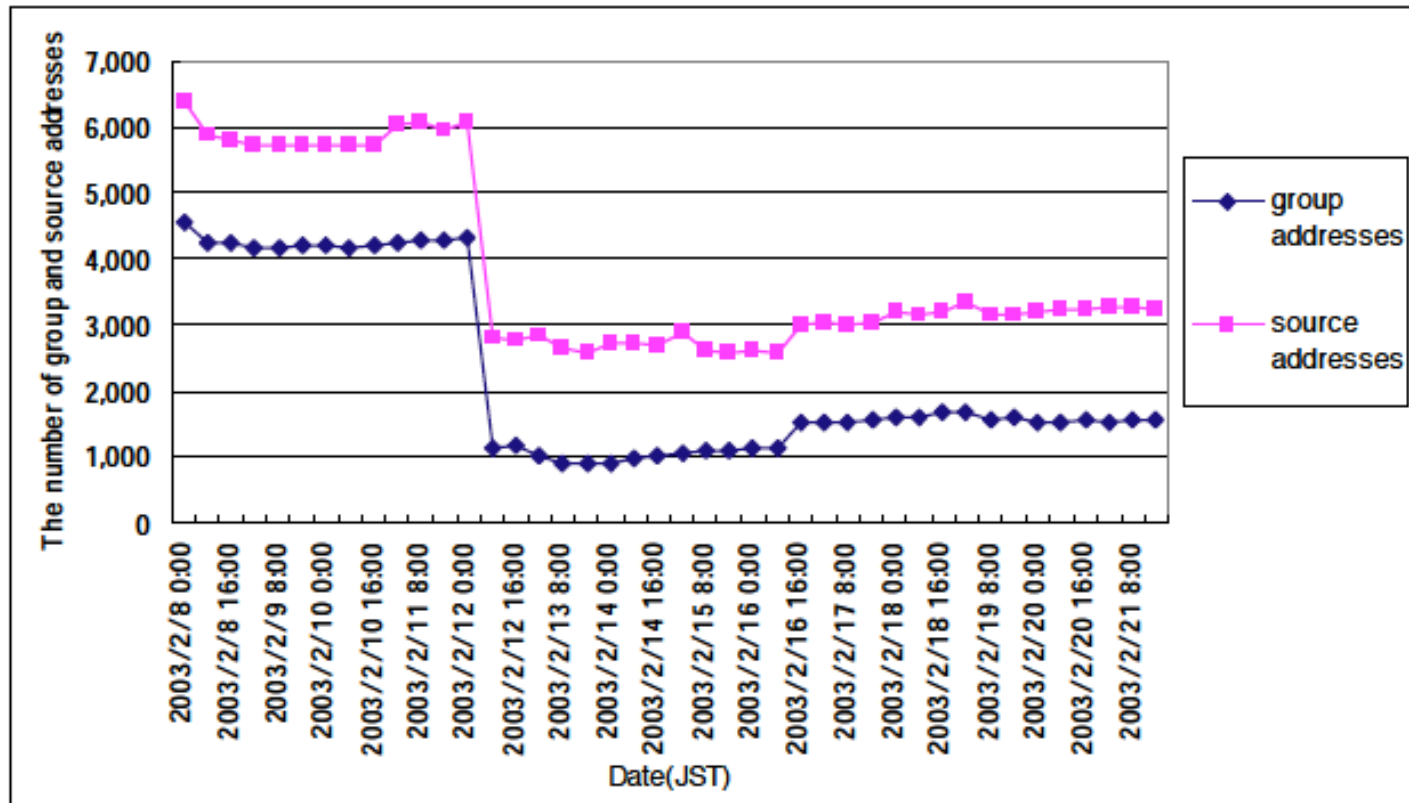


# マルチキャストの展開

- MBGP
  - (マルチキャスト)ネットワークのトポロジ情報/Unicast != Multicast
- PIM-SM
  - マルチキャスト配送木構築
- MSDP
  - マルチキャスト送信者情報

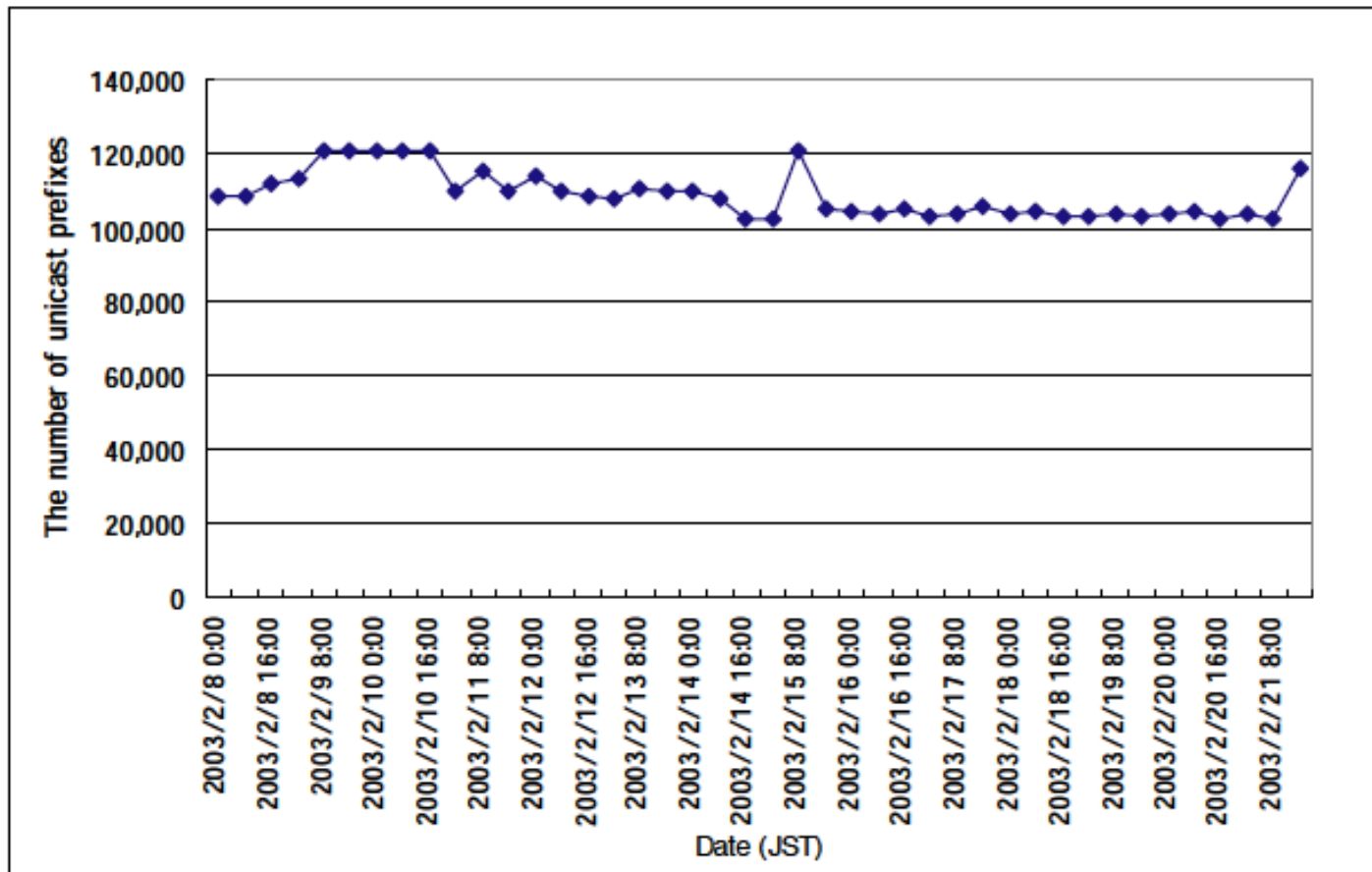
# マルチキャストの展開

## 2.4.2 マルチキャストグループ及び送信者情報数



# マルチキャストの展開

## 2.4.4 ドメイン間でBGPにて交換されている国際ユニキャスト経路数

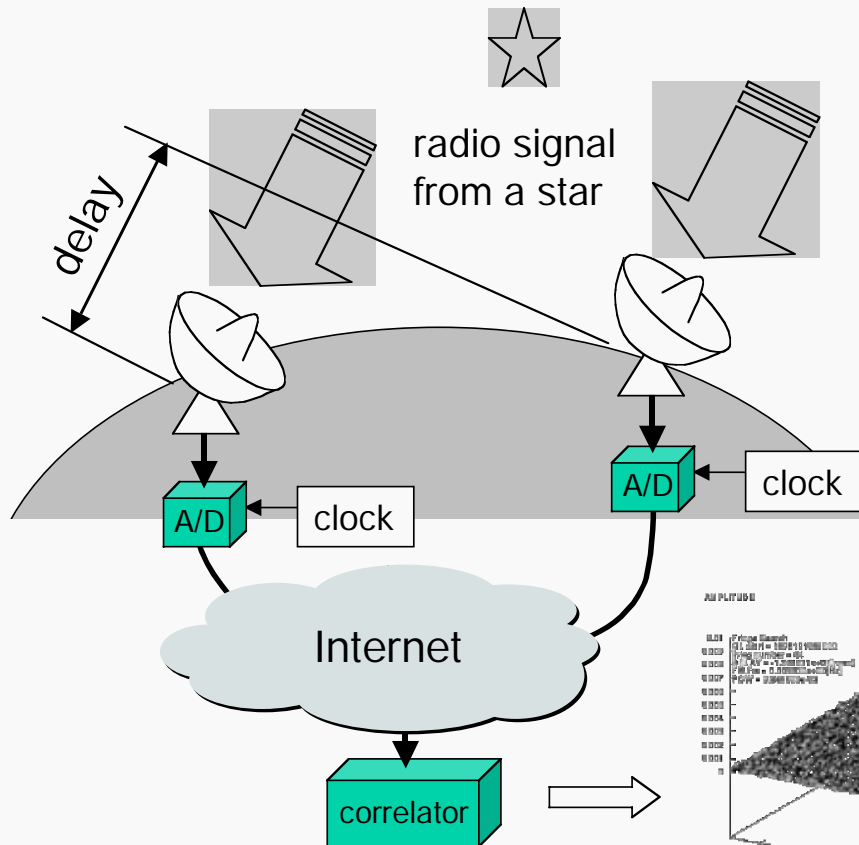


# マルチキャストの展開

- ユニキャスト、マルチキャストネットワークの情報 (BGP/MBGP) から推定したネットワーク規模は
  - 22:1 (prefix 数換算)
  - 6:1 (/24 で正規化後)
- マルチキャストネットワークとして広告されているネットワークの規模はけっして小さくない。
- ソースの数(一般に受信側アプリケーションでも何らかの feed back を返すソースとして動作する) 1,000 - 7,000 とインターネット全体から見ても非常に小さい。

# TCP 性能計測の展開

## VLBI (Very Long Baseline Interferometry)

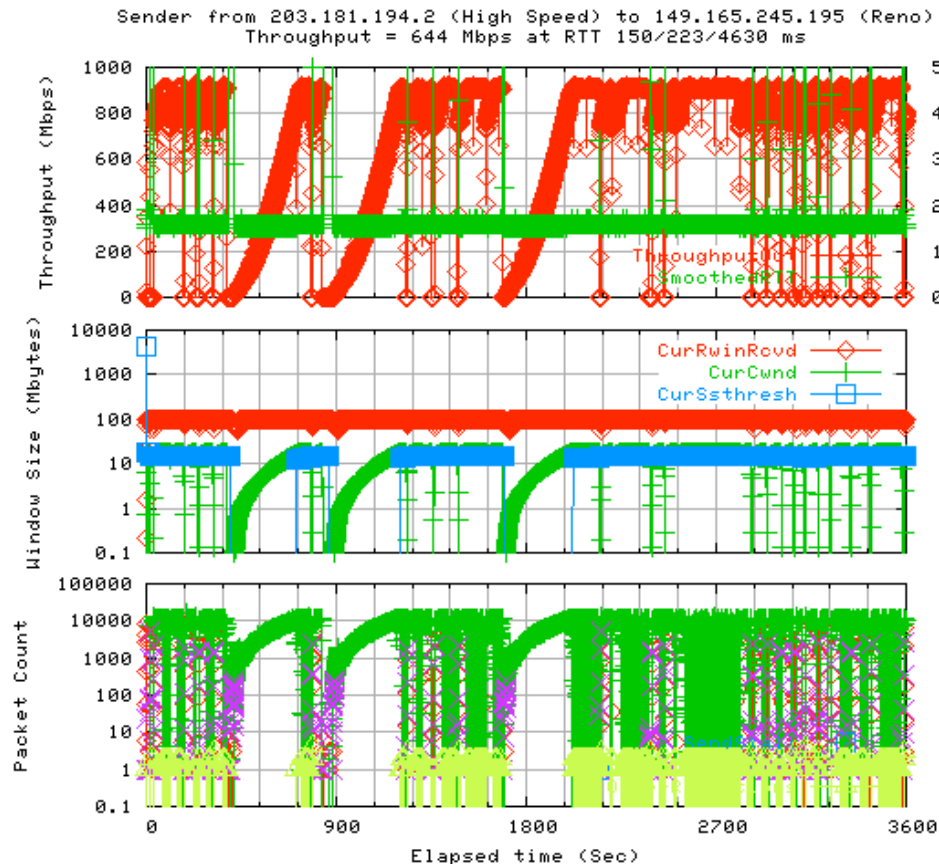


- e-VLBI  
geographically distributed observation, interconnecting radio antennas over the world
- Gigabit / real-time VLBI  
multi-gigabit rate sampling
- ⇒ High Bandwidth –  
Delay Product Network issue

(CRL Kashima Radio Astronomy Applications Group)

# TCP 性能計測の展開

## Advanced TCP Performance Measurement



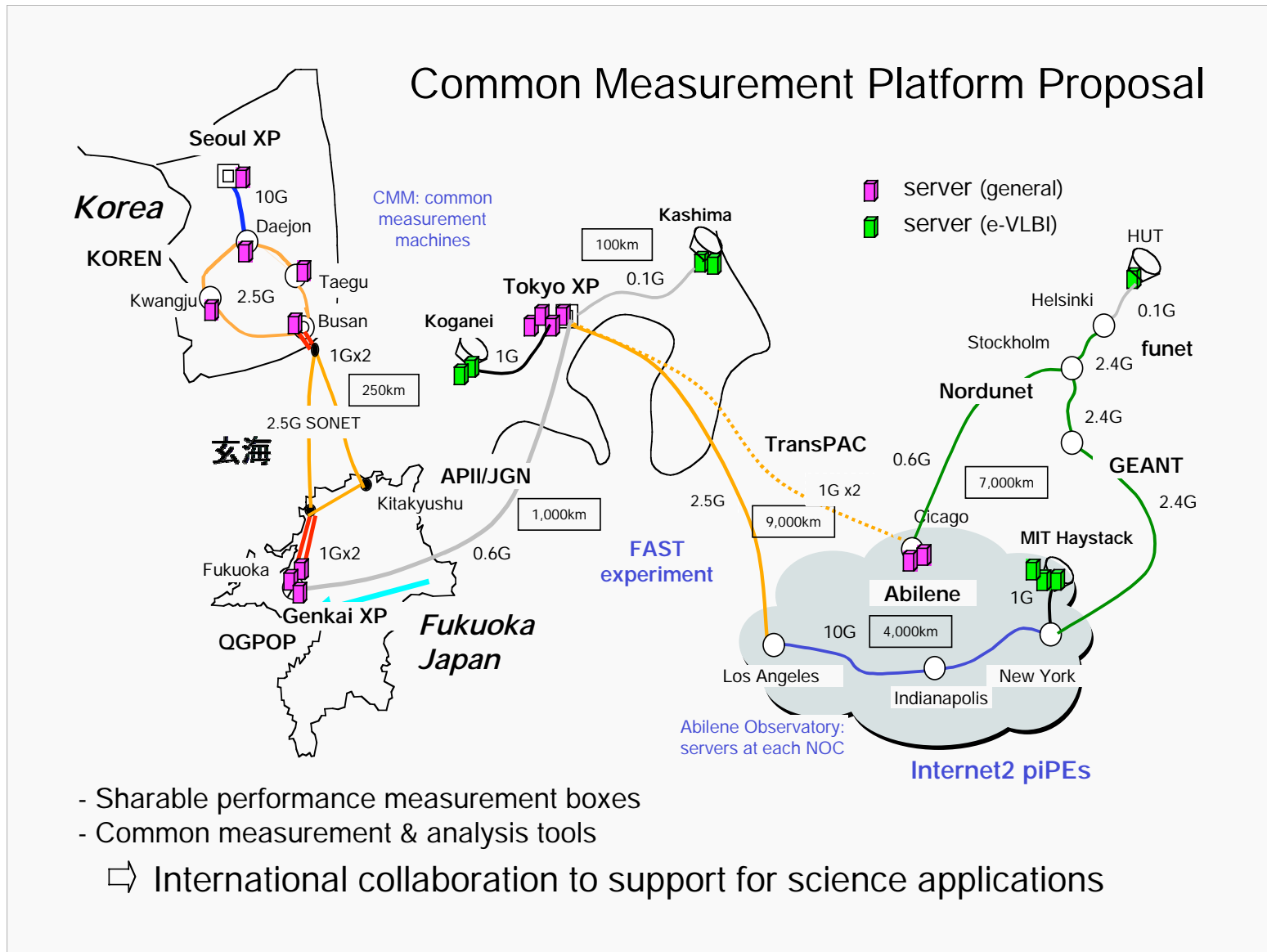
- Measure, analyze and improve end-to-end performance in high bandwidth-delay product networks

- to support for networked science applications

- to help operations in finding a bottleneck

- to evaluate advanced transport protocols (e.g. Tsunami, SABUL, HSTCP, FAST, XCP, ikob)

# TCP 性能計測の展開



# Traffic Analysis for Genkai XP via JGN

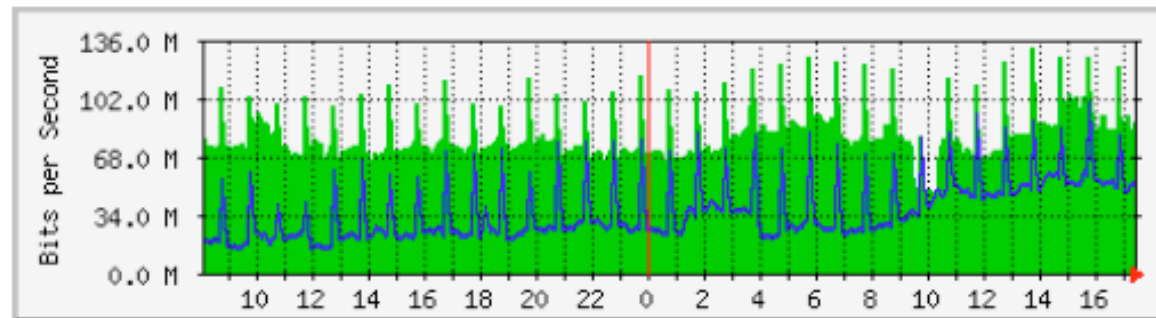
IPv4 PVC (VPI/VCI=4/144)

---

The statistics were last updated **Monday, 1 December 2003**  
**at 17:26**,  
at which time 'tpr3' had been up for **27 days, 6:29:23**.

---

`Daily' Graph (5 Minute Average)



Max 132.8 Mb/s  
In: (21.3%)  
Max 99.7 Mb/s  
Out: (16.0%)

Average 81.1 Mb/s  
In: (13.0%)  
Average 36.6 Mb/s  
Out: (5.9%)

Current 89.3 Mb/s  
In: (14.4%)  
Current 49.7 Mb/s  
Out: (8.0%)

管理職の話はここまで！

# ネットワーク性能は Grid ユーザにとって十分か？

- 極限の end-end ネットワーク性能を得ることは困難を伴う
  - OS, トランスポートスタックの選択
    - Sack, Fast-TCP, Tsunami
  - ネットワークの経路の性能にあわせたチューニング
    - 資源予約？

# Background

- fat-pipe issue will be led by Growing link bandwidth
- 1 2/3 hour is needed to reach max window size for single TCP, in the case of 10Gbps, 100 msec RTT and 1.5KB mss, also unrealistic loss rate (1/5,000,000,000)
  - Some application still require single TCP stream, e.g. network storage.
- Difficult to obtain enough performance under bandwidth difference 10Kbps to 10Gbps, and quality difference from wireless network to fiber.

# Related research

- No additional feature is needed on Router
  - High-speed TCP(S. Floyd et al.) Scalable TCP(T. Kelly)
    - Aggressive window increment when bandwidth may be enough.
  - Fast TCP(Caltech)
    - TCP parameters are tuned by measurement data.
- New feature is needed on Router
  - Quick Start(S. Floyd et al.)
    - Similar BW probe header is sent with TCP SYN. Router compare value and rewrite it if condition is worse. Probe packet is sent every RTT period.
  - XCP (D. Katabi)
    - Router lookup RTT and window size in TCP header, and rewrite it, if router condition is worse.

# Positive Router response

- If end-host can get link information of path,
  - e.g., bandwidth, utilization, and losses,
  - end-host can well tune parameter of TCP, e.g. window increment, slow start threshold, and maximum window size.
- If end-host can detect change of communication path situation,
  - e.g., mobile node step in from PDC network to 802.11a "Hot spot",
  - end-host can switch to appropriate behavior.

# Approach of PR

- Minimize router's work
  - Router does not keep information for each flow or connection, to avoid state explosion.
  - Rate control to avoid congestion should be done by end-host itself.
- Compatible with current TCP
  - not to make impact other original TCP communication, if possible, "TCP friendly" response.

# PR header

- Define new IP option header that router refers and overwrite hop by hop

```
+-----+
= IP header                                     =
+-----+-----+-----+-----+
|Prot.      |req key    |res len.   | res. key   |<-- Router only lookup and overwrite here
+-----+-----+-----+-----+
| TTL       | DATA      |           |           |<-- Router only lookup and overwrite here
+-----+-----+-----+-----+
| TTL       | DATA      |           |           |
+-----+-----+-----+-----+
.
.
+-----+-----+-----+-----+
| TTL       | DATA      |           |           |
+-----+-----+-----+-----+
| TTL       | DATA      |           |           |
+-----+-----+-----+-----+
| TCP, UDP, etc. header and payload          |
+-----+-----+-----+-----+
```

# How to work PR

- Sender transmit packets changing TTL in option header.
- Router compares TTL field in IP header and TTL value in option header.
- If both value is same, router overwrite header field by referred value.
- Receiver send back incoming information with usual response packet.
- After cycle of TTL changing, sender get path link information. Then sender tune parameters of transport control.

# Cost of PR Router

- Router processing cost compared with TCP Quick start
  - Probably not too heavy load recent products.
  - Also router must do other header processing, as TTL decrement.

Proposed way:

```
if (ip.option == this_header_option) then
  if(rp_hdr.data[0].ttl == ip.ttl) then          <- compare
    rp_hdr.data[0].data = data[this_hdr.data[0].id]; <- overwrite
  end
end
```

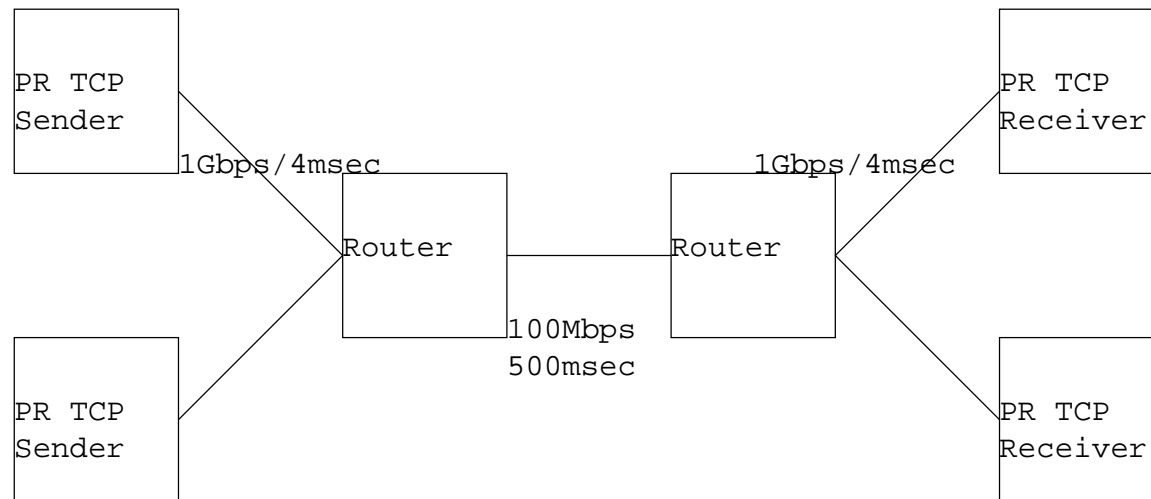
Quick start:

```
if(ip.option == quick_start_option) then
  if(qs_hdr.BW > egress_BW/ingress_BW ) then    <- compare
    qs_hdr.BW = egress_BW/ingress_BW;          <- overwrite
  end
end
```

# PR simulation

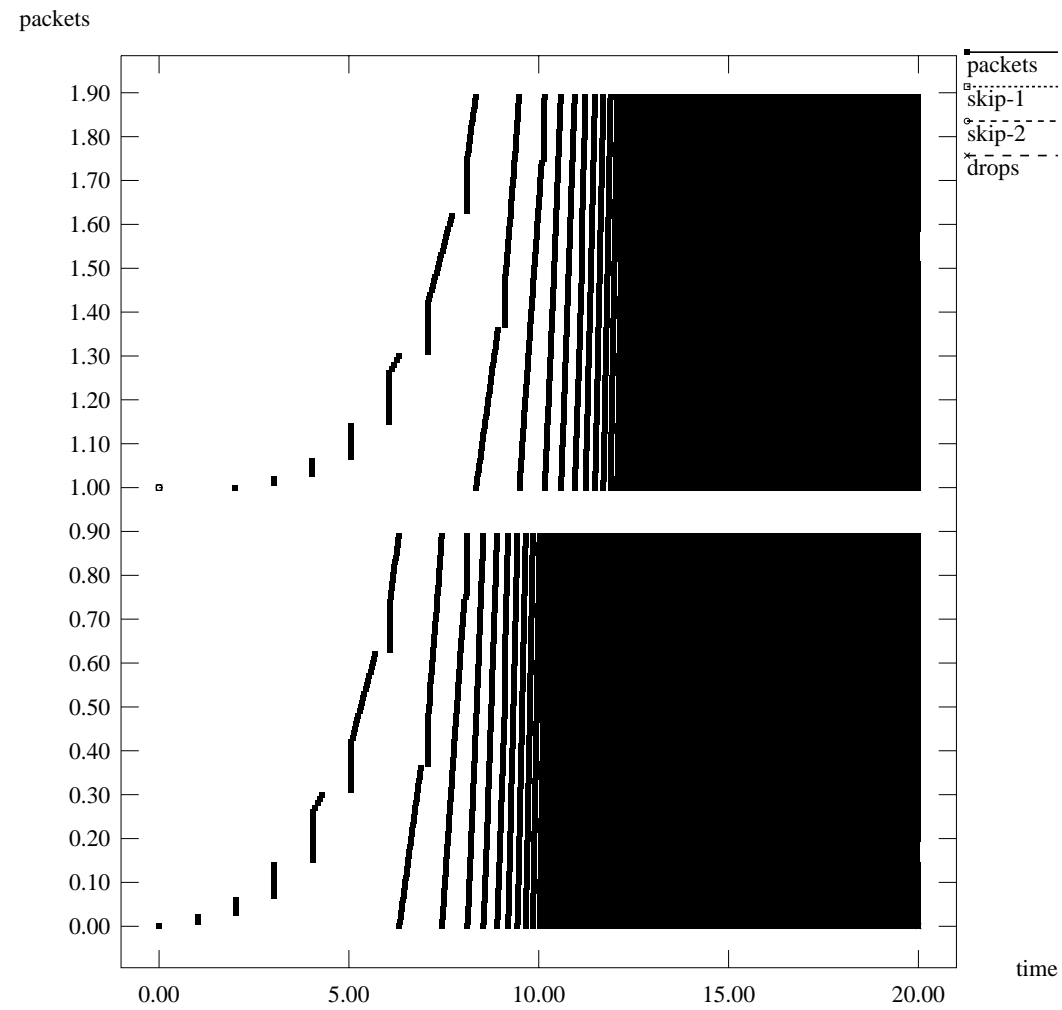
- NS-2.26
  - PR agent and TCP sender receiver
  - lookup bandwidth
    - utilization, loss In/Out, Queue max, link delay
- TCP Reno + RBP
  - $\text{window max} = \text{Bottleneck} \times \text{RTT} \times (1-p)/8 \times \text{MSS}$
  - $\text{ssthresh} = \text{window max}/2$
  - only change above parameters

# A PR result



# A PR result

rbpprr



# Related research

- Although TCP Quick start approach could get information whether every router is compatible or not compatible, it cannot draw path information which part is not compatible, or cannot change object to collect.
- Our way can provide information whole of path compatibility, and also can presume information get from adjacent routers.