

# **Partition Cast - Modelling and Optimizing the Distribution of Large Data Sets in PC Clusters**

*Felix Rauch*

*Christian Kurmann, Thomas M. Stricker*

Laboratory for Computer Systems, ETH Zürich

CoPs project: <http://www.cs.inf.ethz.ch/CoPs/>

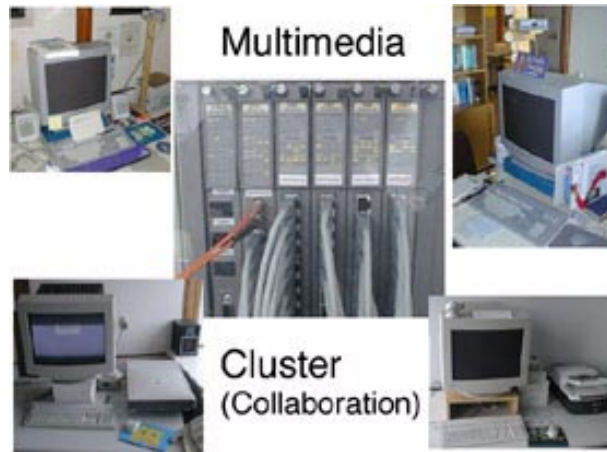


*Eidgenössische  
Technische Hochschule  
Zürich*

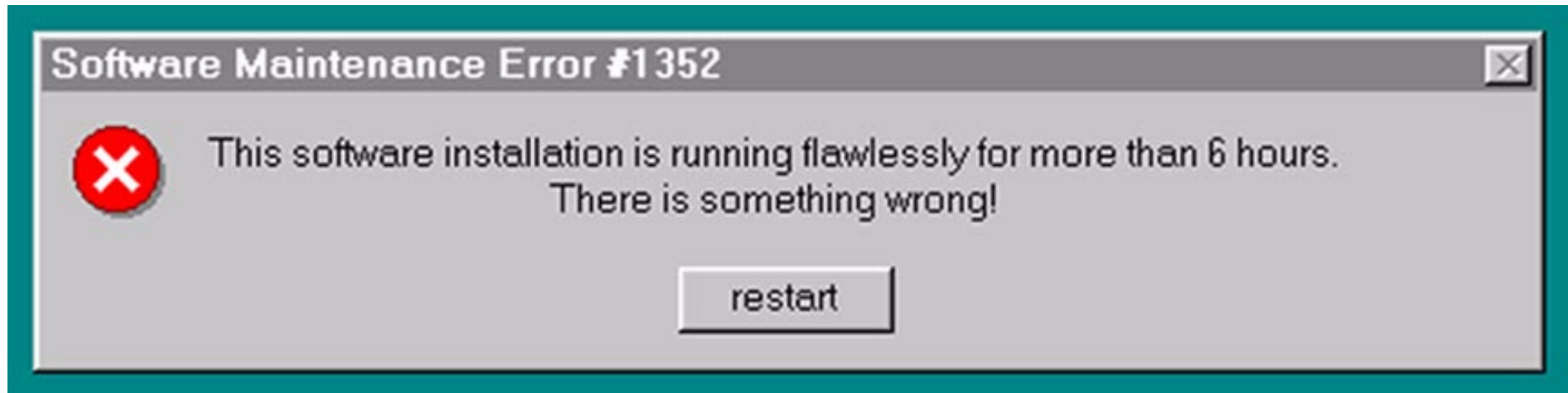
*31. August 2000*

# Clusters of PCs

- Scientific computing (computational grids)
- Enterprise computing (distrib. databases/datamining)
- Corporate computing (multimedia/collaborative work)
- Education and training (classrooms)



# Common Problem



## Maintenance of software installations is hard:

- Different operating systems or applications in Cluster
- Temporary installations: tests, experiments, courses
- Software rejuvenation to combat software rotting process

**Manual Install:** days, **Network Installs:** hrs, **Cloning:** min

# Partition Cast (cloning)

**Fast replication of entire system installations (OS image, application, data) on clusters is helpful**

- How to do ultra fast data distribution in clusters?

**Essential tradeoffs:**

- **What network** is needed? Giga/Switches/Hubs
- **Protocol family**? (multicast, broadcast, unicast)
- **Compressed or raw** data?
- **Best logical topology** for distribution path?

# Overview

- Network topologies and embedding
- Related work
- Analytical model for partition cast
- Implemented tools for partition cast
- Evaluation of alternative topologies
- Model vs. measurement
- Conclusion

# Network Topologies

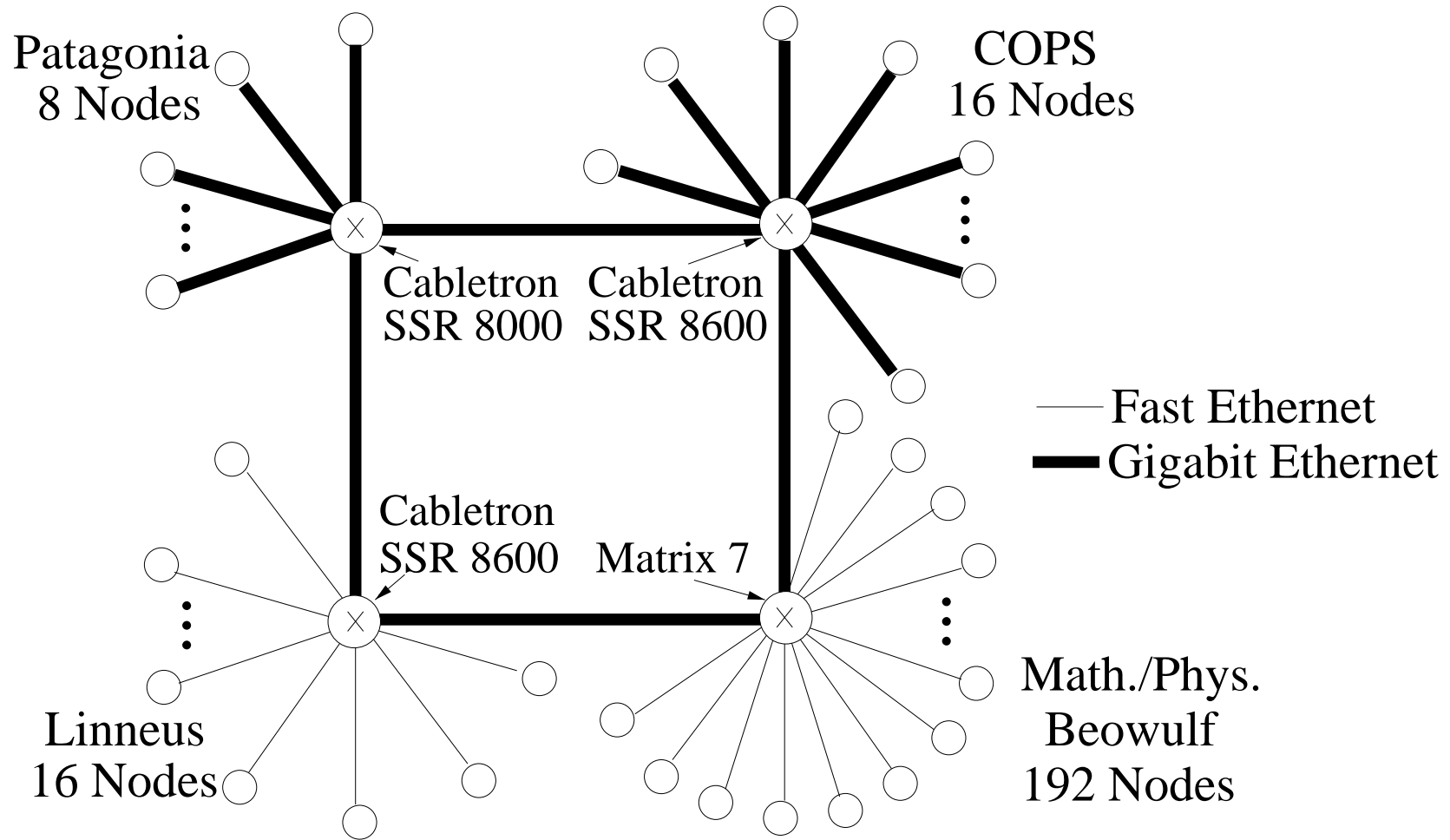
## Given:

- **Physical network topology**
- **Resource constraints**  
(maximal throughput over **links** or trough **nodes**)

## Wanted:

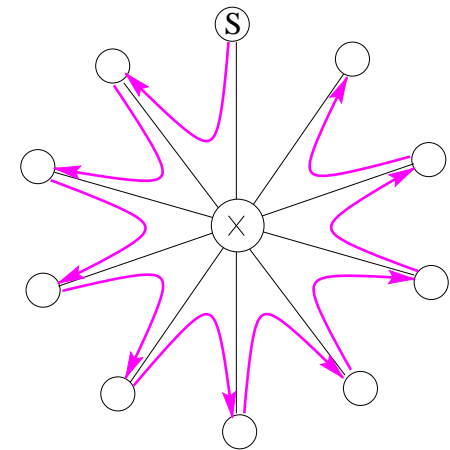
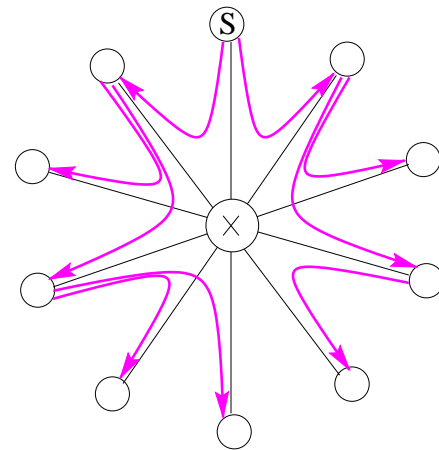
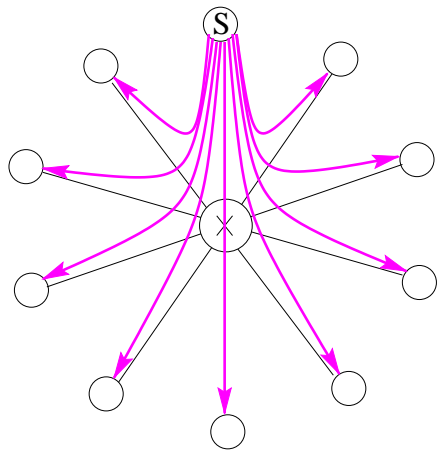
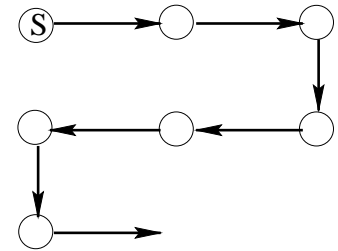
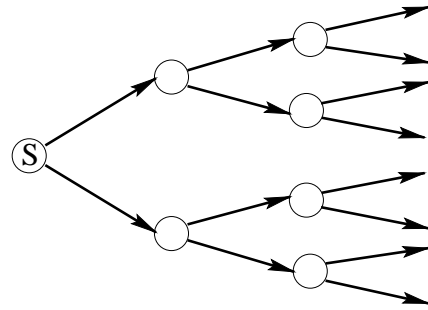
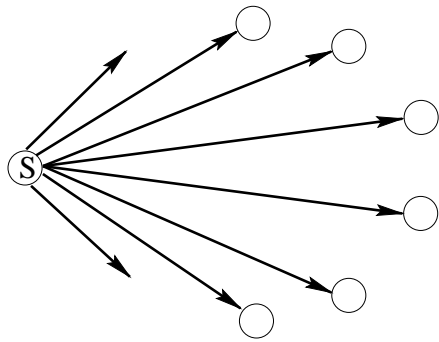
- Best **logical network topology** for data distribution.
- Best **embedding** of logical network into physical network?
- Limit on **throughput** for distribution of big data sets  
(partition cast)

# Physical Network



- Graph given by cables, nodes and switches

# Logical Network



- Spanning tree, embedded into physical network

# Previous and Related Work

- Protocols and tools for the distribution of data to large number of clients  
[Kotsopoulos and Cooperstock, **USENIX 1996**]
- Model is based on ideas for throughput-oriented memory system performance for MPP computers  
[Stricker and Gross, **ISCA95**]
- High speed multicast leads to great variation in perceived bandwidth, complex to implement and quite resource intensive. High speeds seem impossible.  
[Rauch, masters thesis, **ETHZ 97**]

# Simple Model of Partion Cast

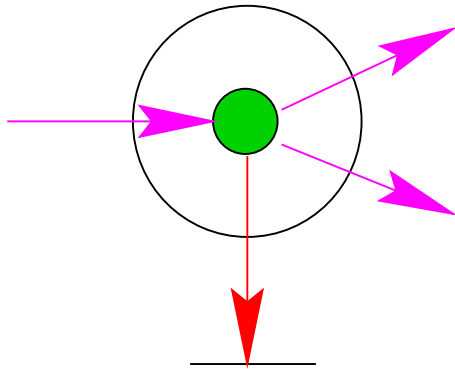
## Definitions:

- **Node types**
- **Capacity constraints**
- **Algorithm for evaluation** of model

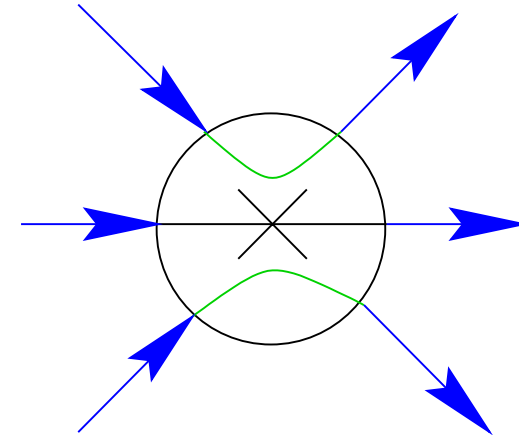
## Example:

- Heterogenous network: Gigabit / Fast Ethernet

# Node Types



Active node



Passive node

- Active node: Participates in partition cast, can duplicate and store stream
- Passive node: Can neither duplicate nor store data, passes one or more streams between active nodes

# Capacity Constraints

- Reliable transfer promise
- Fair sharing of links

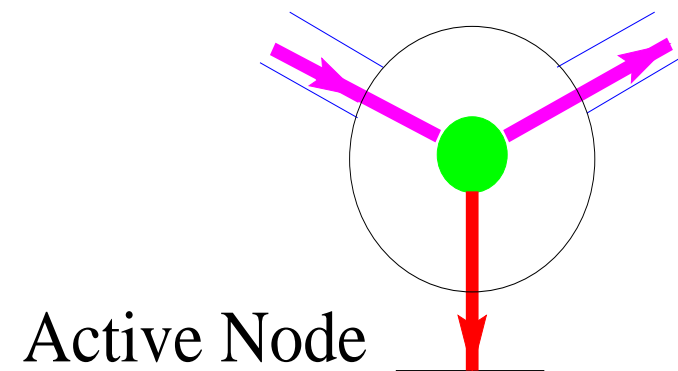
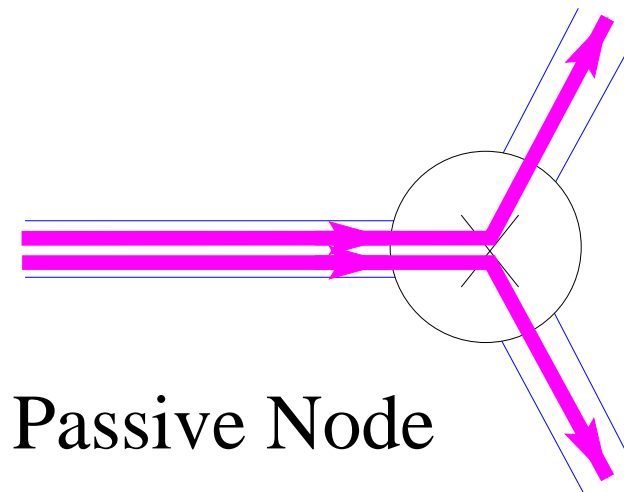
- **Edge** capacity

*Link 125 MB/s, 2 logical channels  $\rightarrow$   $<62$  MB/s*

- **Node** capacity

*Switches 30 MB/s, 3 Streams  $\rightarrow$   $<10$  MB/s*

Examples:

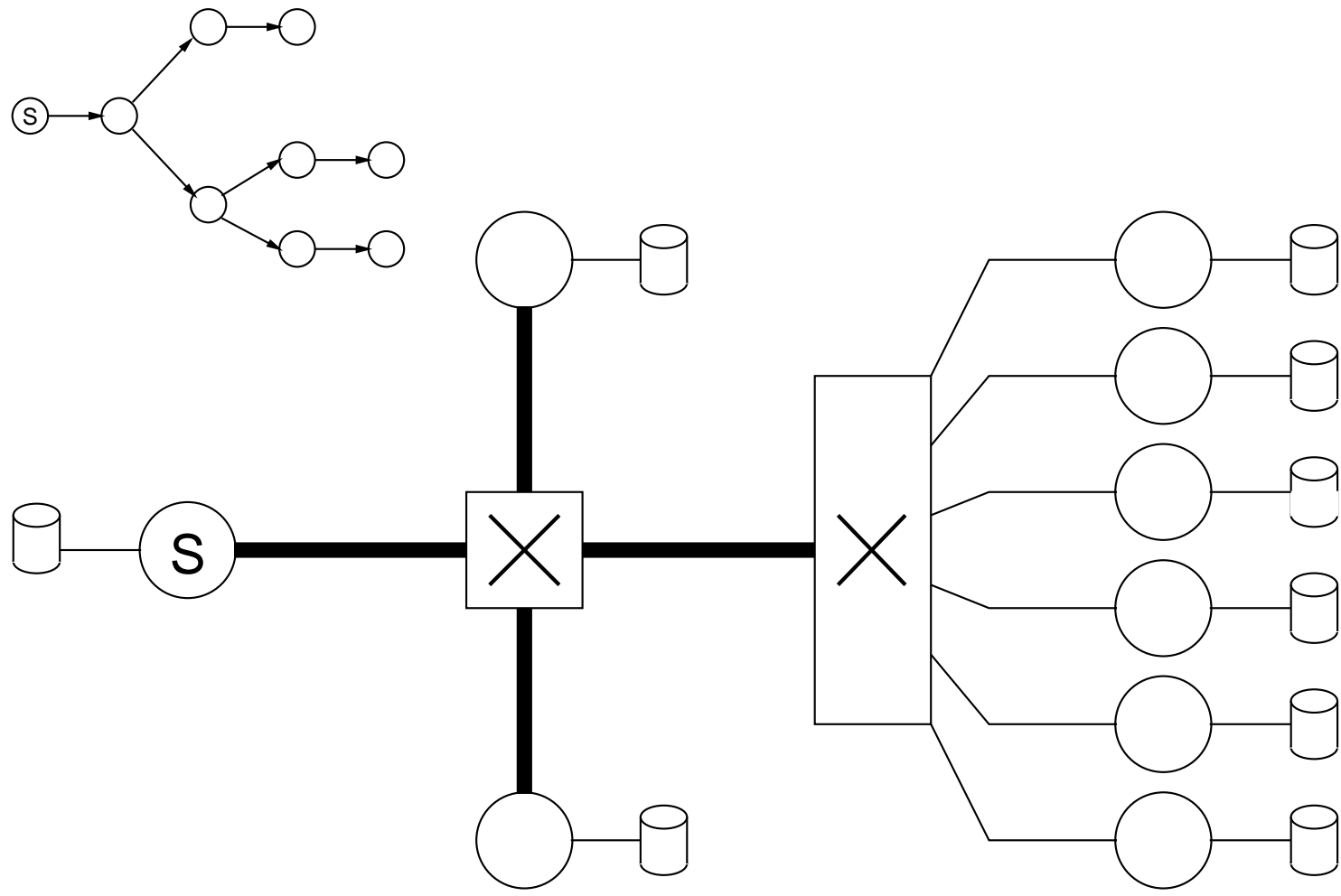


# Model Algorithm (Constraint Satisfaction)

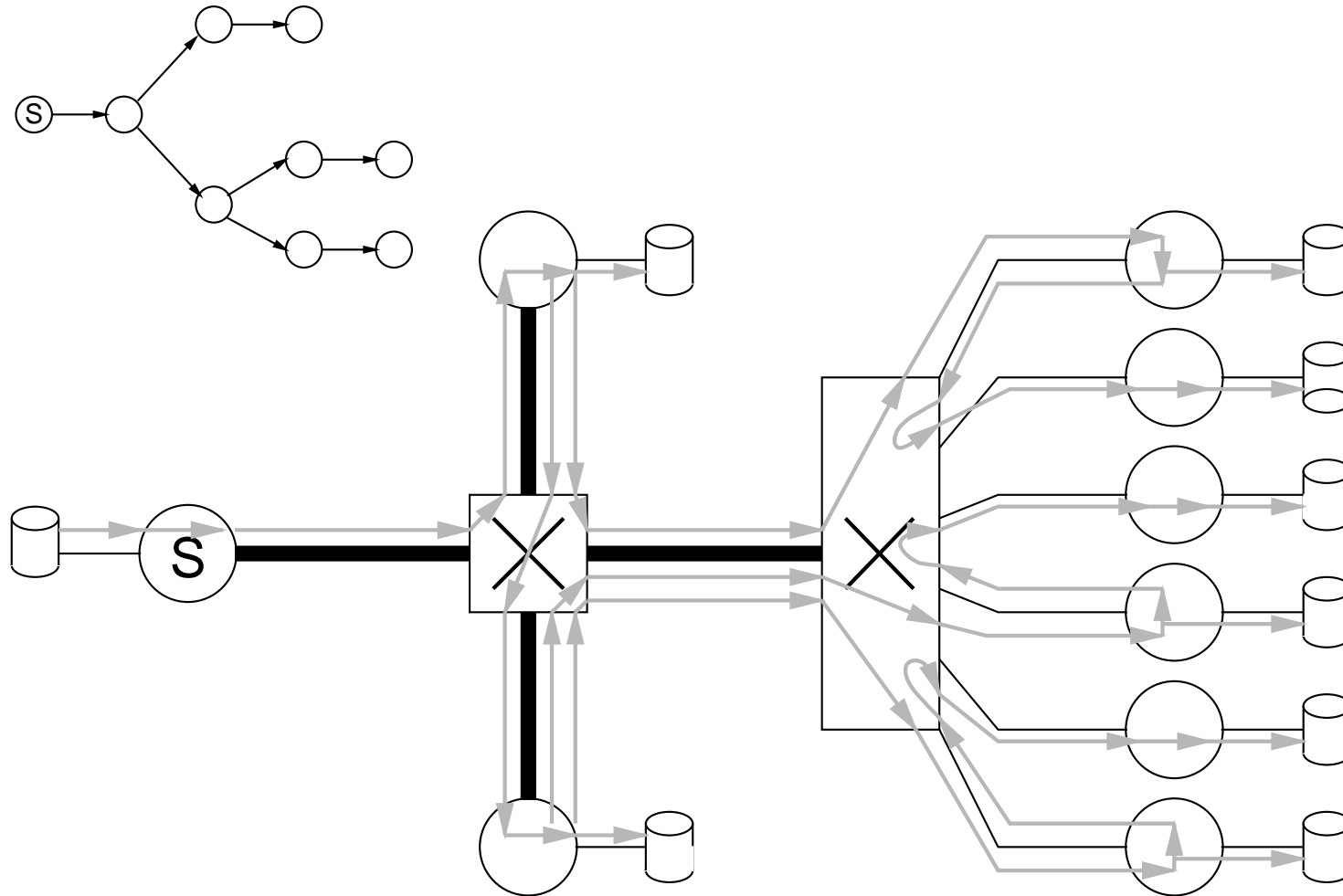
*Algorithm* "evaluate basic model"

- 1 **Choose** logical network
- 2 **Embed** into given physical network
- 3 *For all edges*  
    **Post bandwidth limitations** due to edge congestions
- 4 *For all nodes*  
    **Post bandwidth limitations** due to node congestions
- 5 *Over all posted limitations*  
    **Find minimum bandwidth**

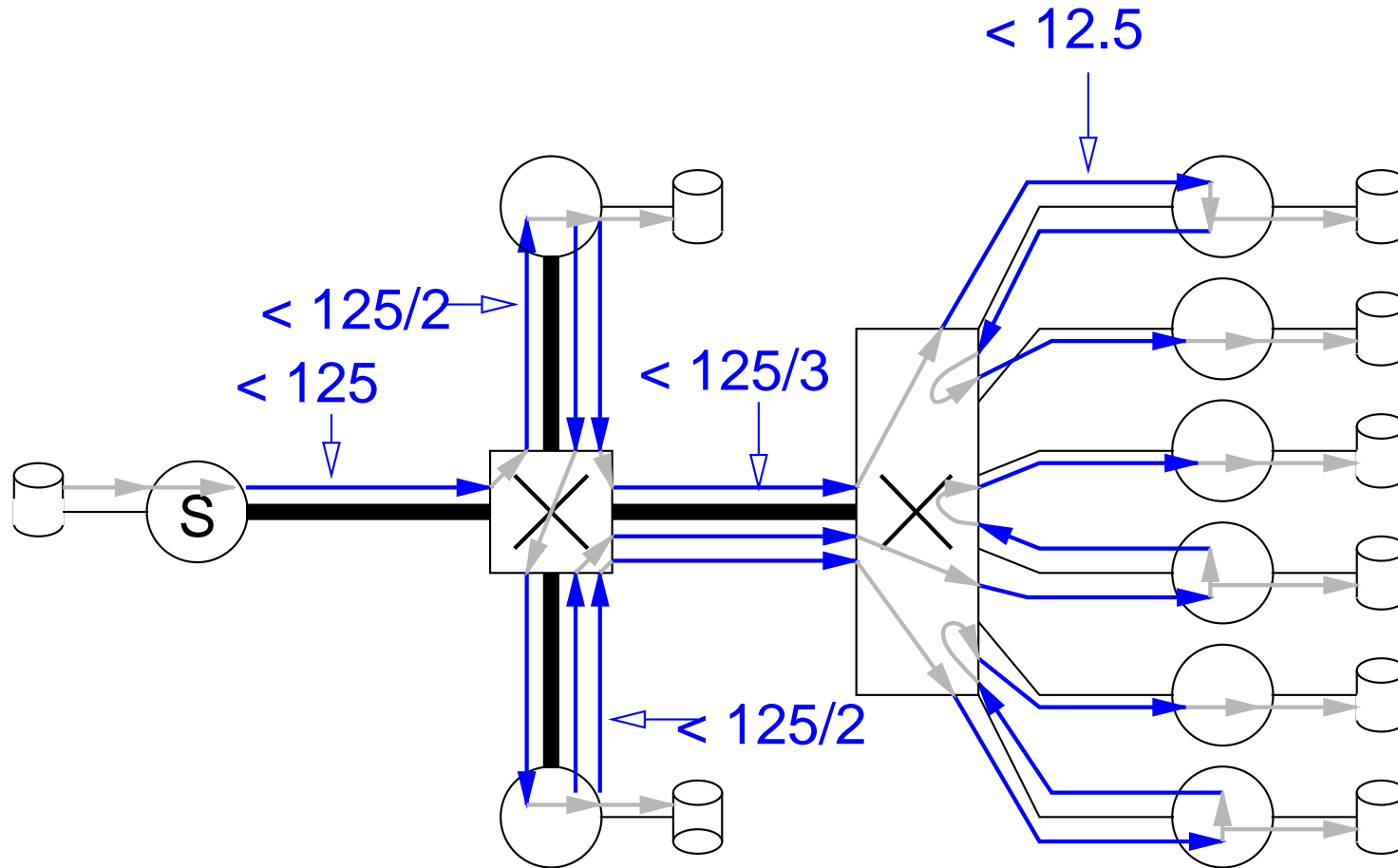
# Example Network



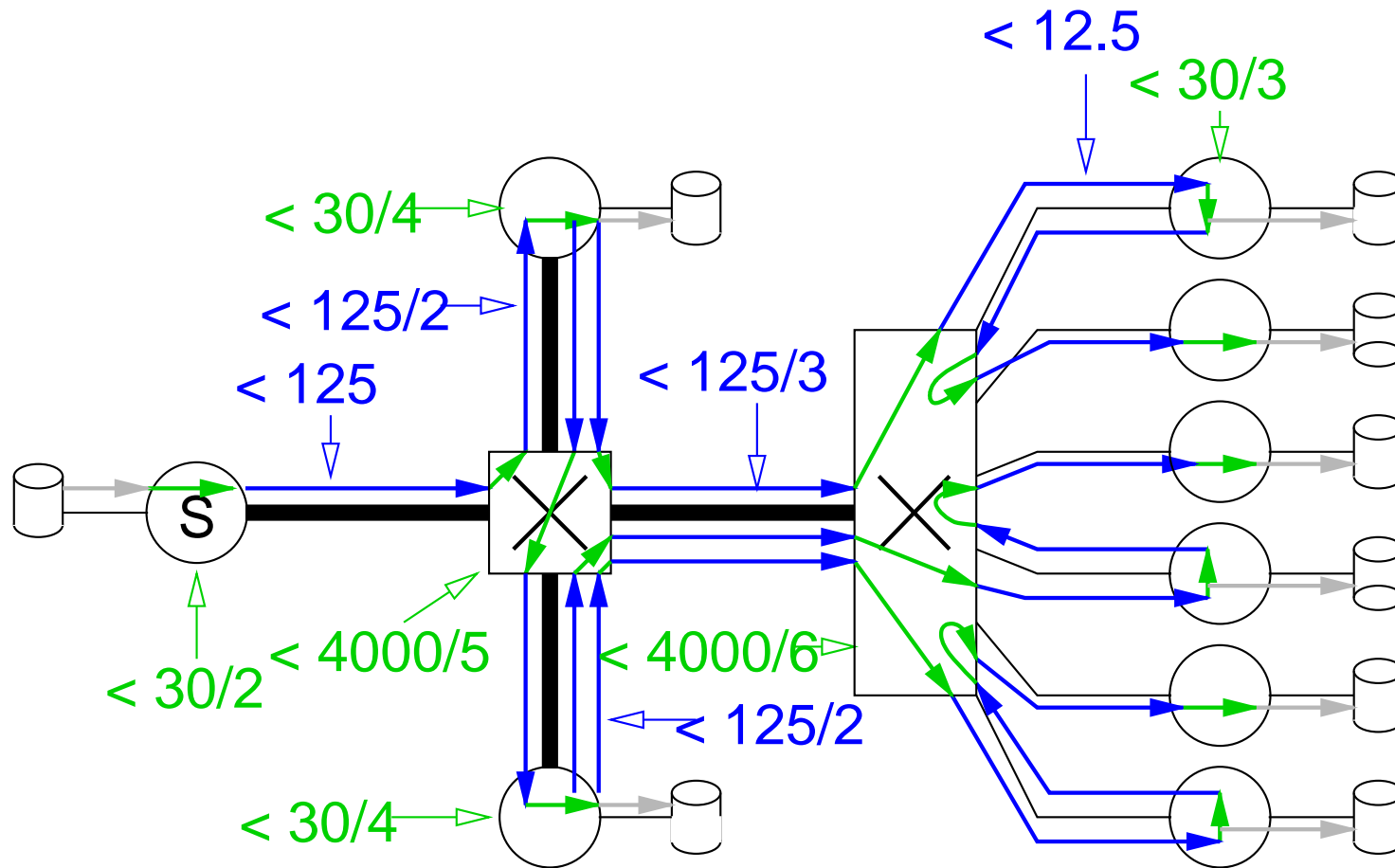
# Example Network



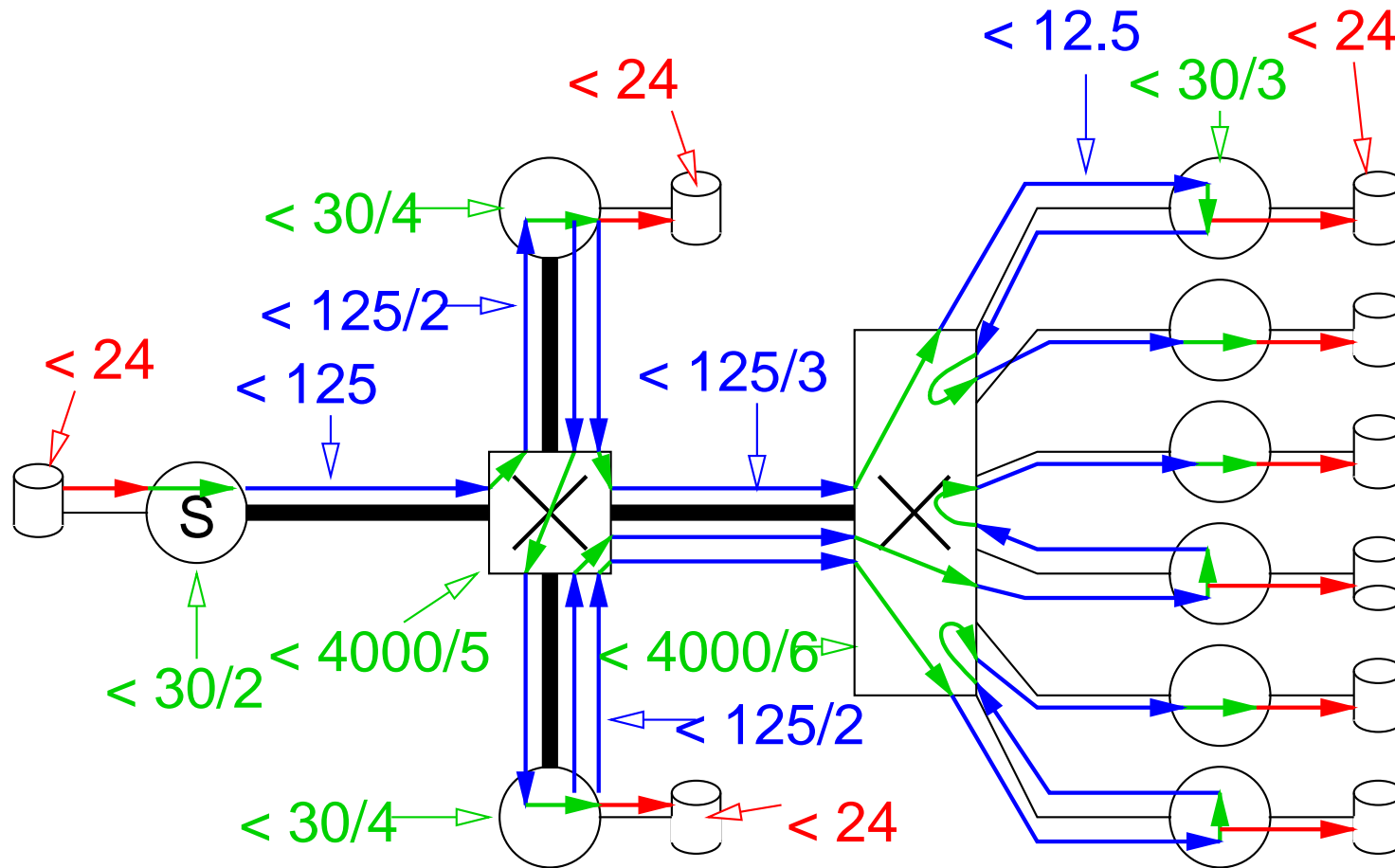
# Example Network



# Example Network



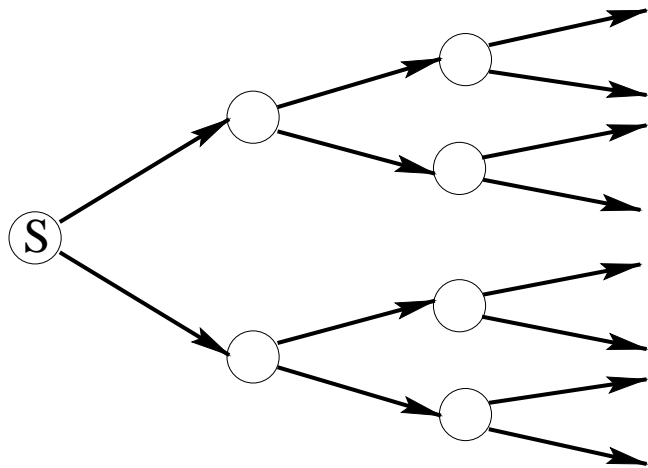
# Example Network



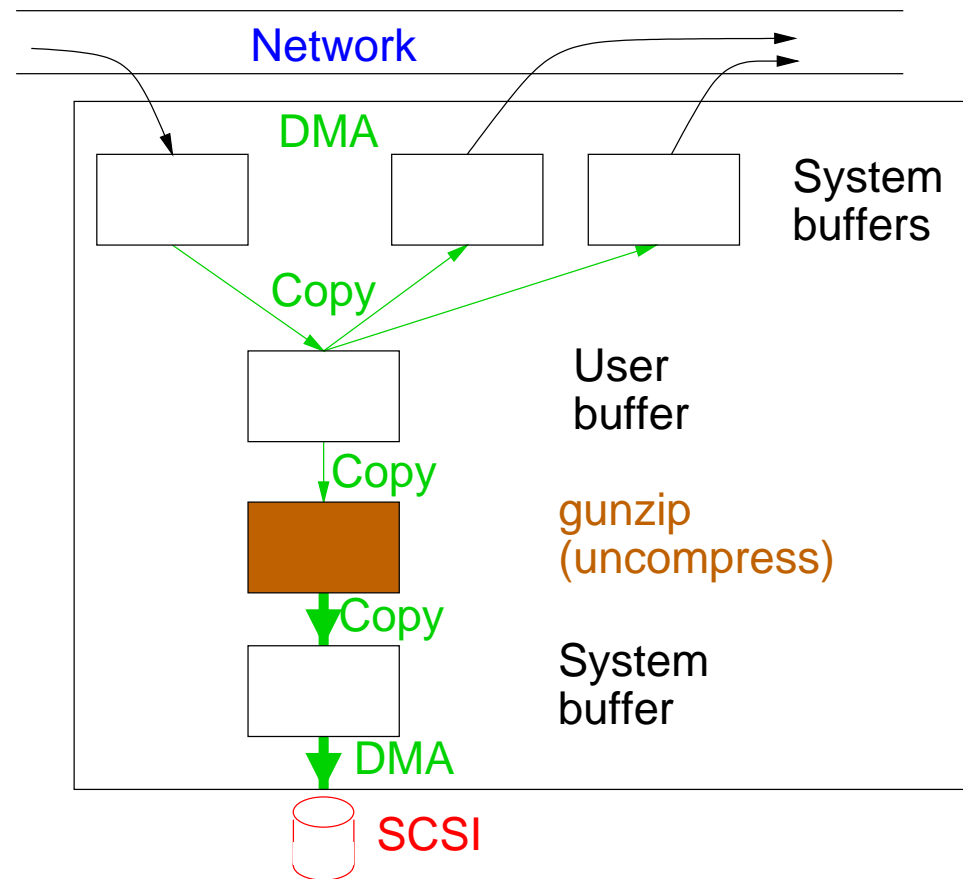
# Detailed Model of Active Nodes

- In the simple model active nodes were black boxes
- Detailed model would allow accurate predictions of achievable data stream bandwidths
- Requires detailed knowledge of:
  - Flows of node-internal data streams
  - Limits of involved subsystems
  - Complexity of handling and coordinating data streams and subsystems

# Detailed Example: Data-Streams



Logical Topology



Data streams within active node

# Limitations in Active Nodes

- **Link capacity**

Gigabit Ethernet: 125 MB/s

Fast Ethernet: 12.5 MB/s

- **Disk system**

Seagate Cheetah SCSI harddisk: 24 MB/s

- **I/O bus capacity**

Current 32 bit PCI bus: 132 MB/s

- **CPU utilization**

Processing power required for each stream, depending on speed and complexity of handling

# Detailed Example of an Active Node

- Modelling switching capacity: Binary spanning-tree topology with Fast Ethernet and compression:

const c:  
compression  
factor  
e.g. c=2

$$\frac{b}{c} < 12.5 \text{ MB/s} \quad \textit{link receive}$$

$$\frac{2b}{c} < 12.5 \text{ MB/s} \quad \textit{link send}$$

b: bandwidth

$$b < 24 \text{ MB/s} \quad \textit{SCSI disk}$$

$$\frac{3b}{c} + b < 132 \text{ MB/s} \quad \textit{I/O, PCI}$$

$$\frac{8b}{c} + 3b < 180 \text{ MB/s} \quad \textit{Memory}$$

$$\left( \frac{3}{45c} + \frac{1}{80} + \frac{4}{90c} + \frac{1}{90} + \frac{c}{9} \right) b < 1 \text{ (100\%)} \quad \textit{CPU}$$

**Solve equations for b -> node can handle 5.25 MB/s**

# Implementation (tools for partition cast)

- **dd/NFS**, built-in OS function and network file system based on UDP/IP - simple - permits star topology only
- **Dolly**, small application for streaming with cloning based on TCP/IP - reliable data streaming



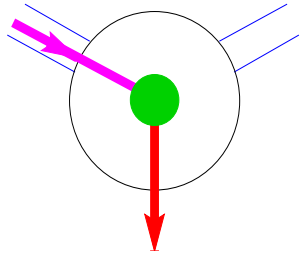
## Dolly

for reliable data casting  
on all spanning trees

- star (n-ary)
- 2-ary, 3-ary
- chain (un-ary)

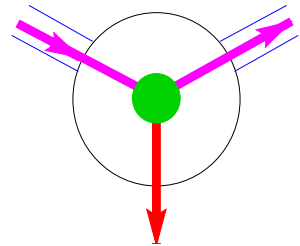
# Active Nodes with Dolly

- Simple receiver for star topologies



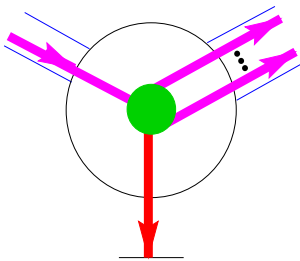
Simple receiver

- Advanced cloning node for multi drop chains



Multi-drop receiver

- Node cloning streams for general spanning trees



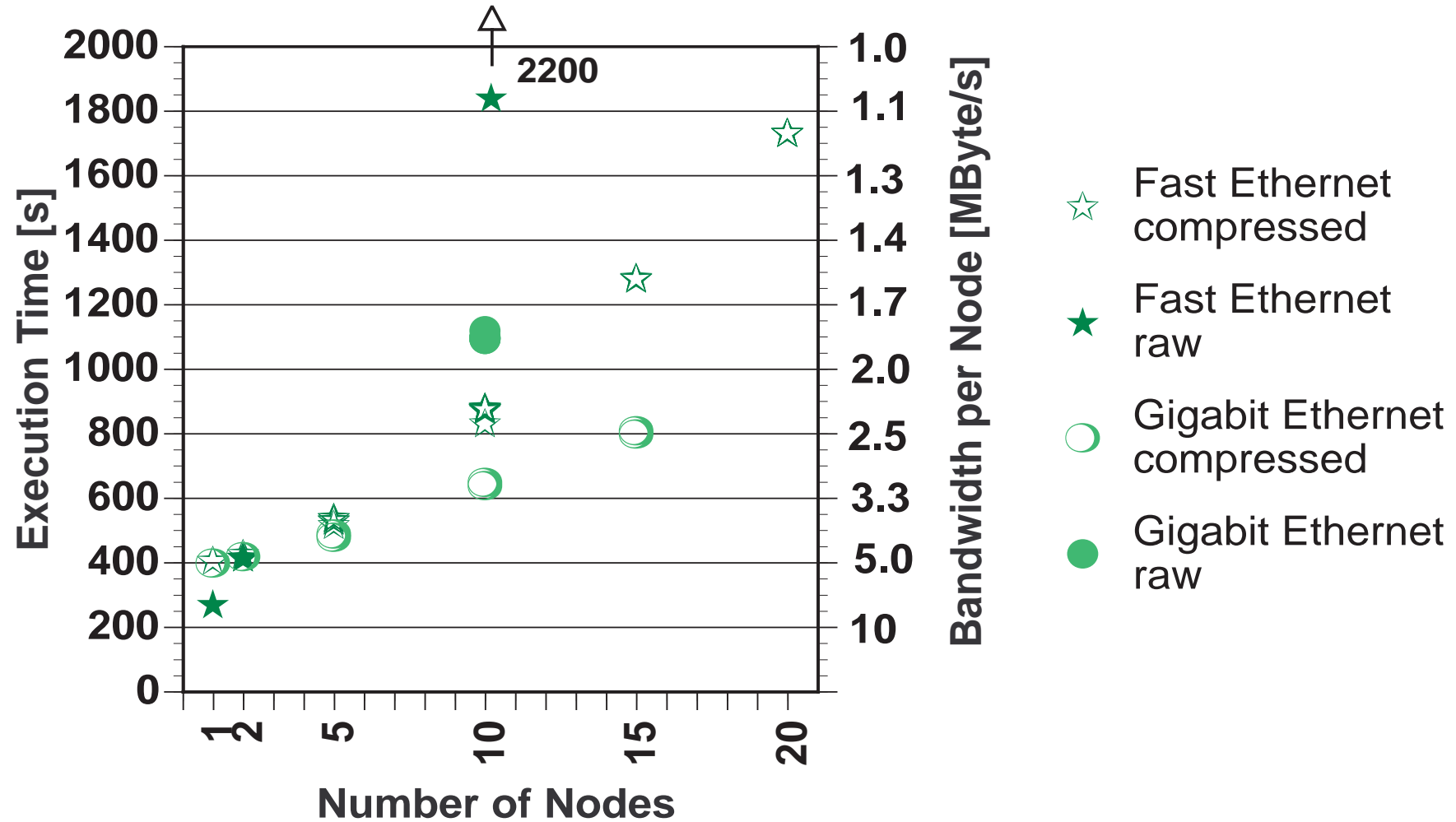
Active node cloning streams

# Experimental Evaluation

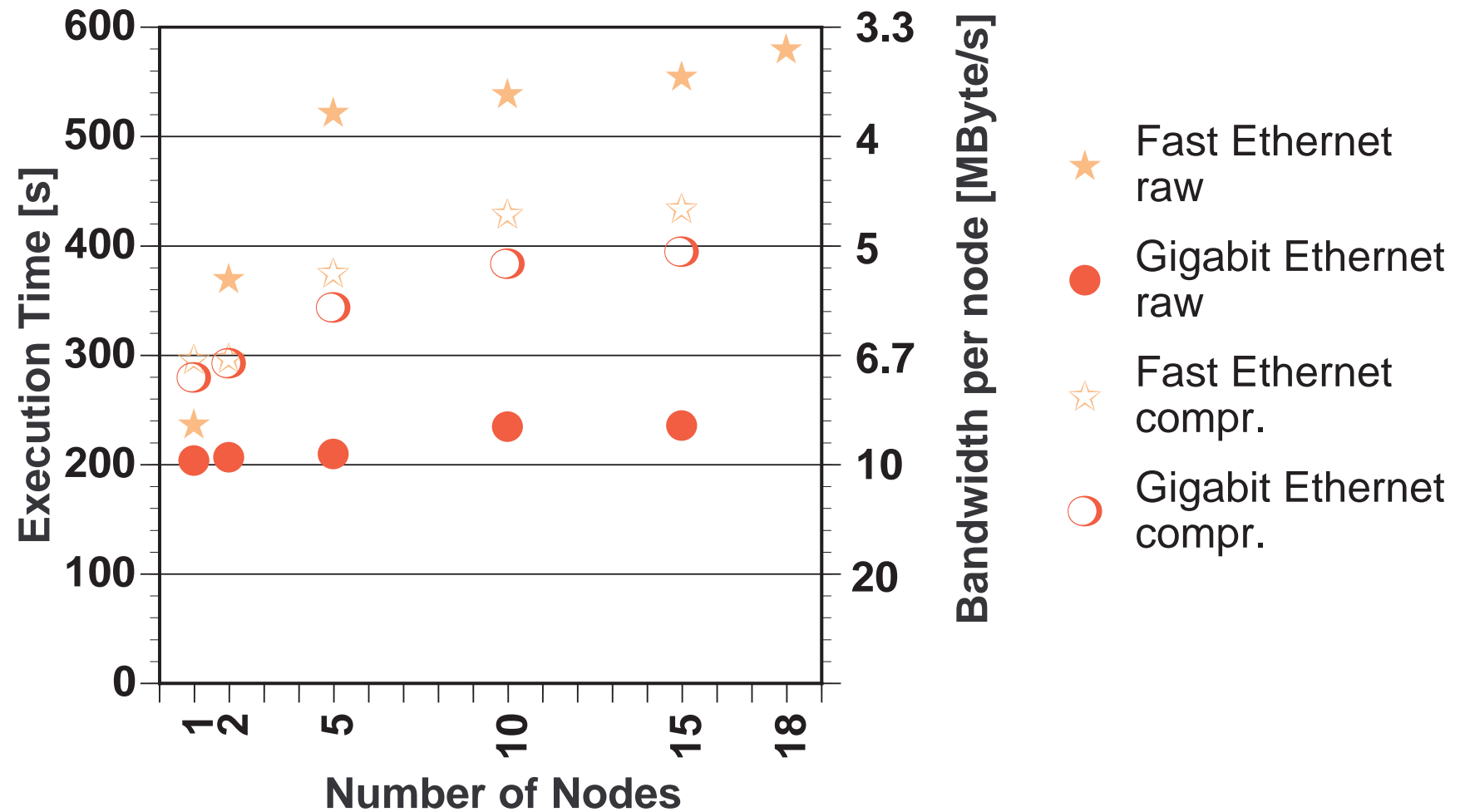
- Topologies:
  - **Star**
  - **3-ary spanning tree**
  - **Multi-drop chain**
- Fast Ethernet / Gigabit Ethernet
- Compressed / Uncompressed Images

All experiments: Distribute 2 GByte to 1..15 clients

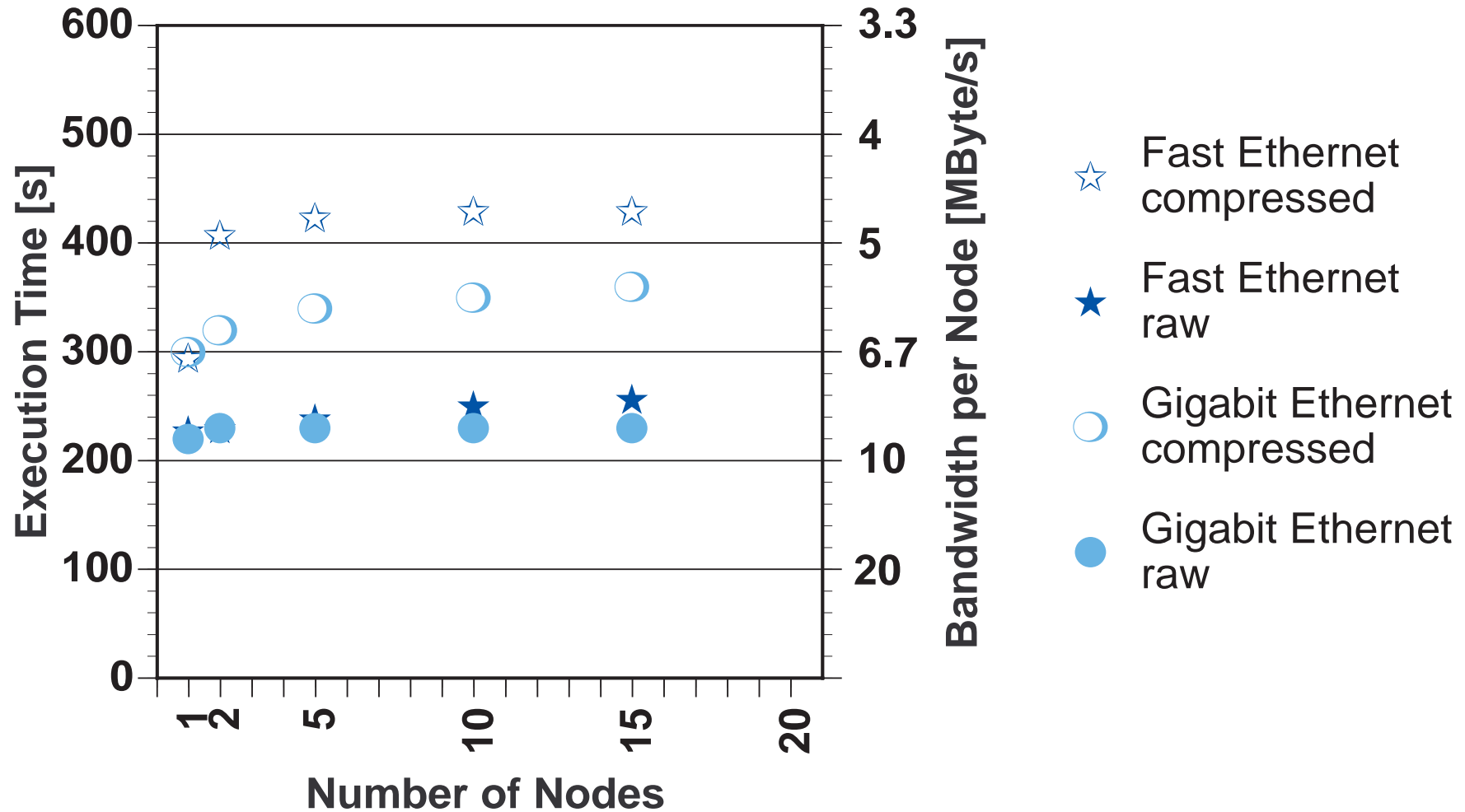
# Star topology (Standard NFS)



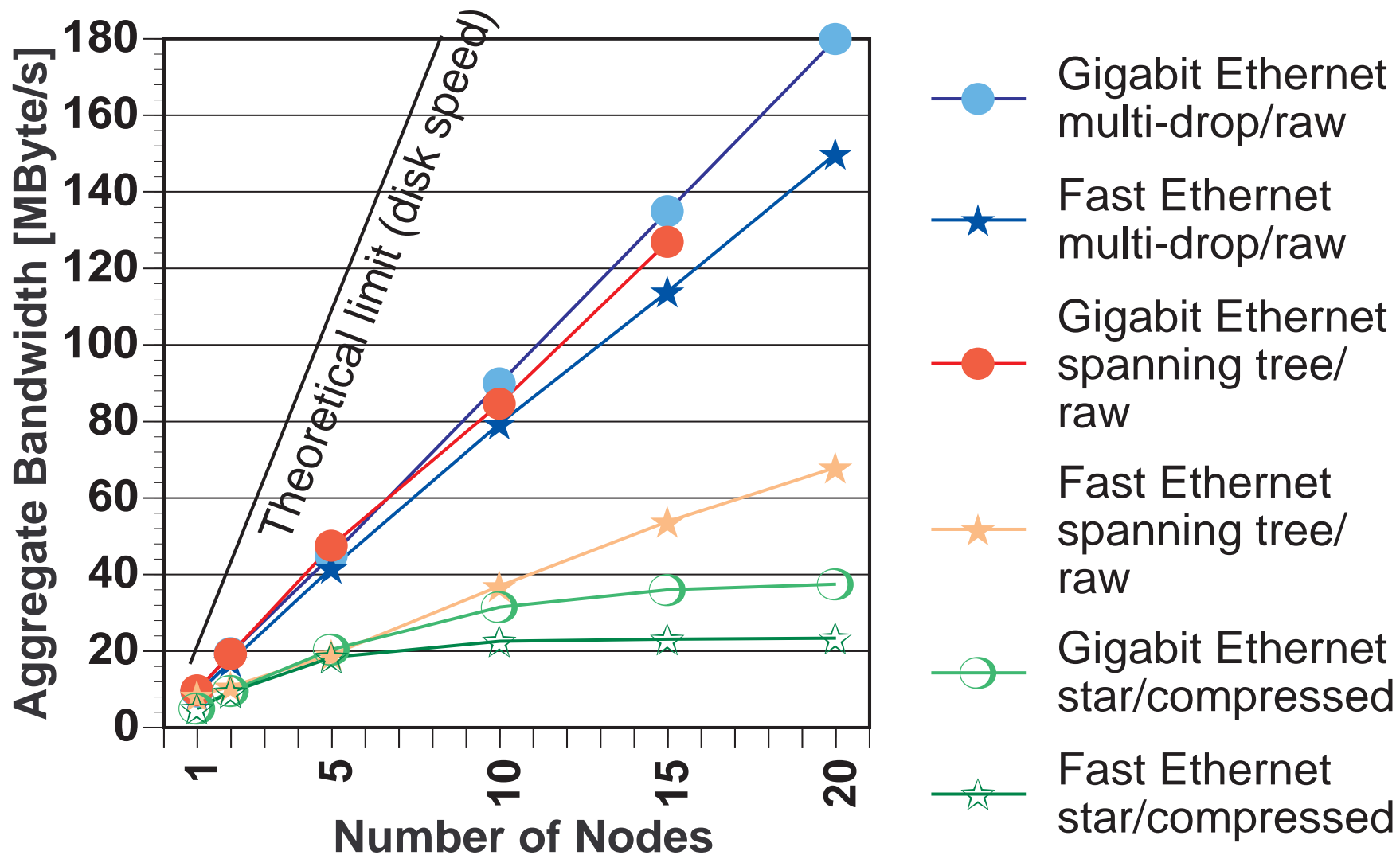
# 3-Tree (Dolly)



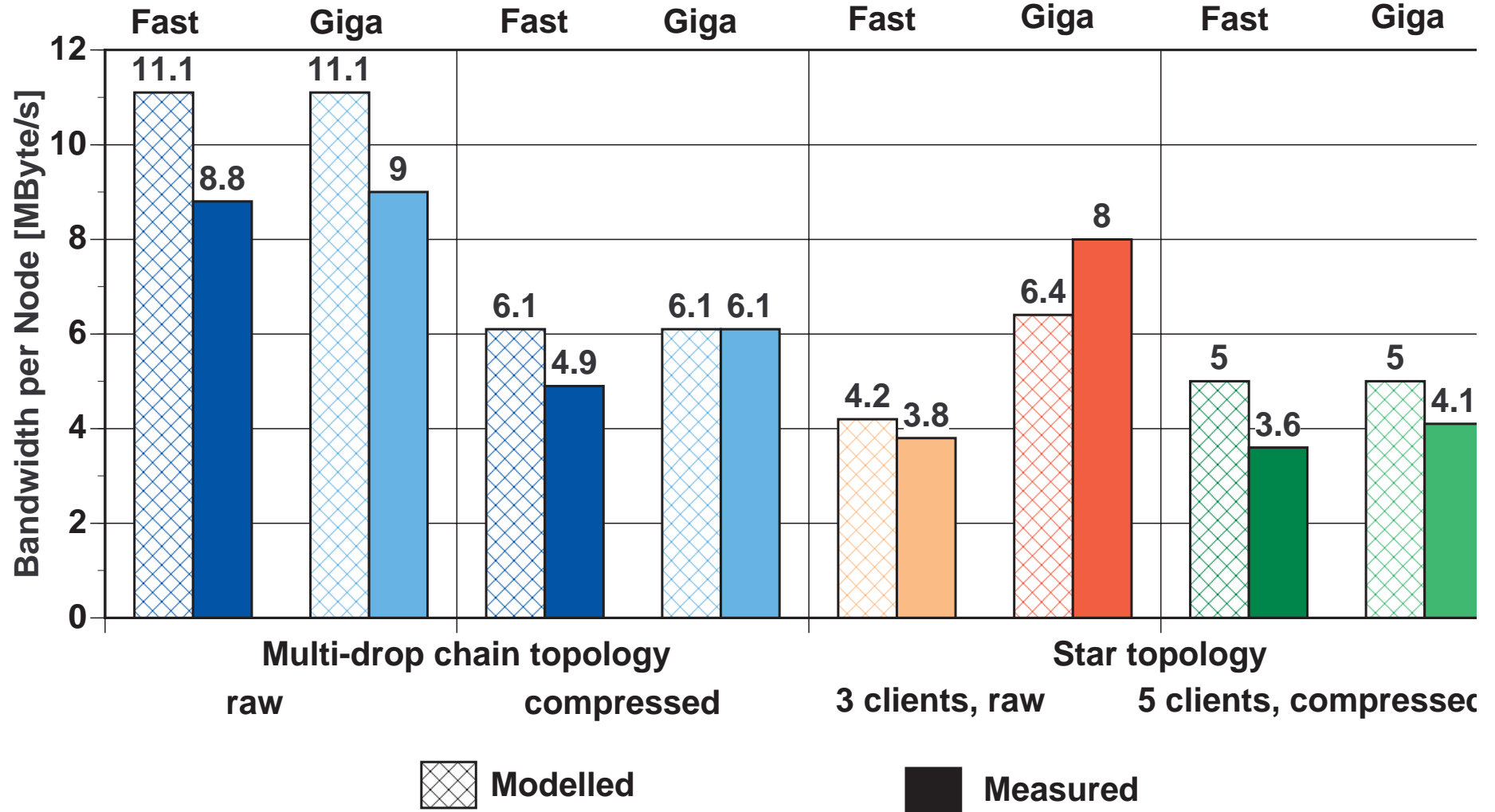
# Multi-Drop Chain (Dolly)



# Scalability



# Predictions and Measurements



# Conclusions

- A **simple model** captures network topology and node congestion
- An **extended model** also captures the utilisation of basic resources in nodes and switches.
- Optimal configurations can be derived from our model.
- For most physical networks a linear **multi-drop chain** is better than any other spanning tree configuration for distributing large data sets.
- **Dolly** - our simple tool - transfers an entire 2 GB Windows NT partition to *24 workstations in less than 5 minutes*, with a sustained transfer rate of 9 MB/s per node

# Questions/Discussion?



## Our Project

**CoPs - Cluster of PCs**

Lab for Computersystems

ETH Zürich, Switzerland

Dolly is available for download under the GNU general public license (source code included).

<http://www.cs.inf.ethz.ch/CoPs/>