End of an Era ...
969798990001 02
Eve of a New Millennium

Super Computer Communications

Ralph Niederberger
Forschungszentrum Jülich GmbH
R.Niederberger@fz-juelich.de

Introduction

- Introduction
- GTB West
  - Goals, Projects, Timeframes and Configuration
  - Super Computer Impediments and Solutions
- Status of Cray Super Computer Communications
- Future Tests
- Summary

Introduction

- New kinds of Microprocessors and expansion of internal storage lead to new kinds of supercomputing systems solving best different kinds of problems.
- Two mostly known types of supercomputers are massively parallel systems and vector systems.
- A new kind of supercomputer is the Metacomputer.
- A Metacomputer distributes an application onto 2 or more equal or distinct machines which are coupled dynamically via an external network.
- This distribution may be done by quality (functional distribution) or by quantity.

GTB - West

Project sponsored by BMBF and DFN with financial participation of the project partners

Partners:

Research Center Jülich GmbH	http://www.fz-juelich.de
GMD - Nat. Res. Center for Inform. Technology	http://www.gmd.de
Deutsches Klimarechenzentrum	http://www.dkrz.de
Alfred Wegener Inst. for Polar & Marine Res.	http://www.awi.de
Pallas GmbH	http://www.pallas.de
o.tel.o	http://www.o-tel-o.de

Runtime: Aug, 1st 1997 - Jan, 31th 2000

More Info: http://www.fz-juelich.de/gigabit

GTB West - Goals

- Demonstrate the usefulness of high speed wide-area communication networks for scientific computing
- Engage in selected applications which are known to need very high communication bandwidth
- Major objective:
  - coupling of architecturally different supercomputers
    i.e. vector computers and massively parallel computers

  fi to build a new kind of metacomputer
- strengthen the know how in
  - high speed computer communications,
  - metacomputing in LAN and WAN environments
  - coupling of the super computer centers in Germany

Impediments

Current problem:

Communication throughput within and between supercomputers differs extremly

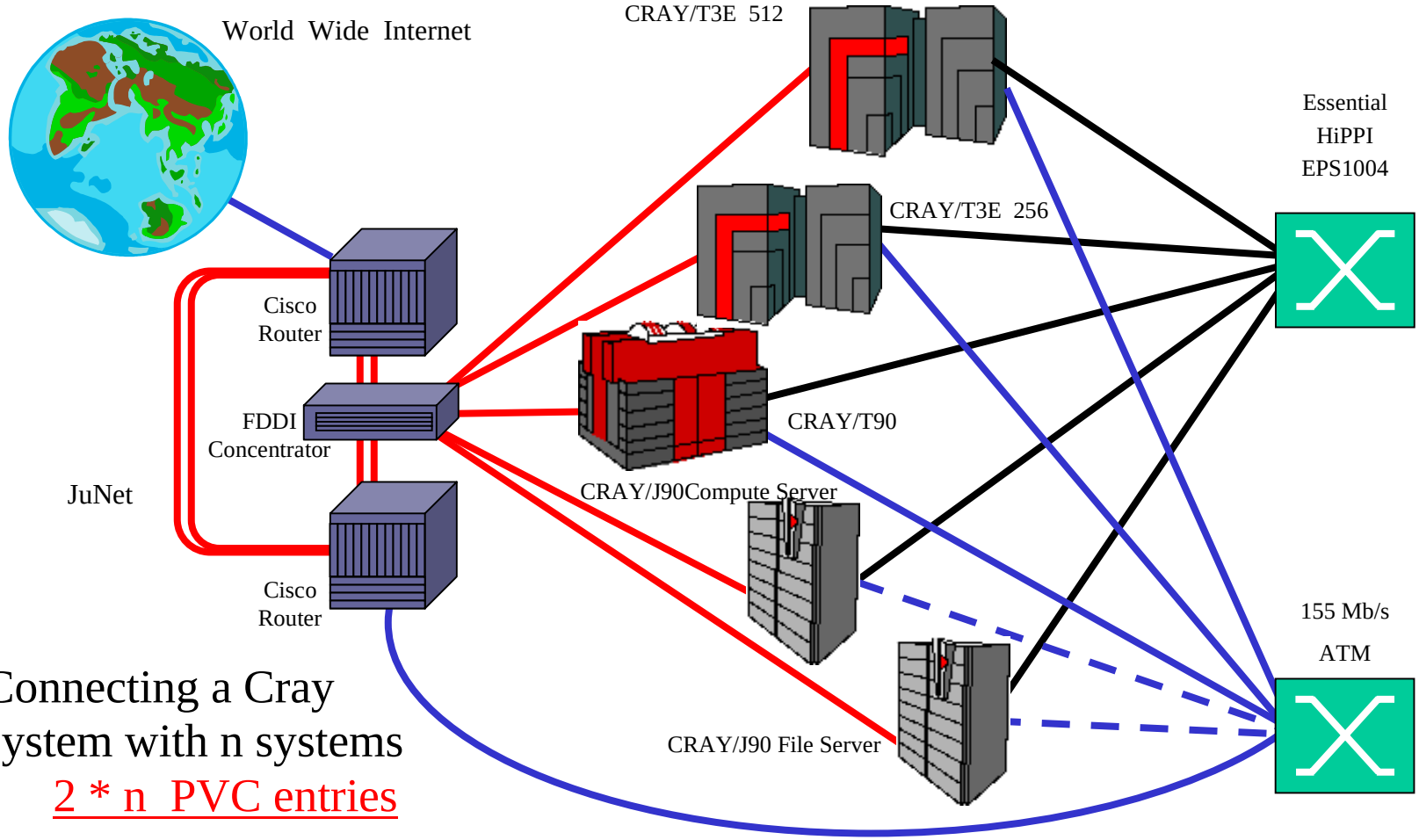
Example:

Cray/T3E with internal communication throughput of 500 MB/s bidirectional into three dimensions (3D torus)

High speed external connections:

~~(Fast-) Ethernet (10-100 Mb/s)~~, ~~FDDI (100 Mb/s)~~ ,
HiPPI (800 Mb/s-1600 Mb/s), ~~Super HiPPI (6400 Mb/s)~~,
~~ATM 155 Mb/s, 622 Mb/s - 2.4 Gb/s~~, ~~Gigabit Ethernet (1Gb/s)~~,

Cray Systems Network Environment



Connecting a Cray system with n systems

2 * n PVC entries

High speed communication

Alternatives communicating between CRAY/T3E and IBM/SP2

- rawHiPPI (800 Mb/s)
  - HiPPI Tunneling (622 Mb/s, currently MTU 9180)
  - HiPPI Sonet Extender (currently 155 Mb/s or 932 Mb/s)
- TCP/IP via HiPPI (622 Mb/s, currently MTU 9180 because of routing)
- nativeATM (155 Mb/s, 622 Mb/s) **(Hardware ?, Software ?)**
- TCP/IP via ATM (155 Mb/s, 622 Mb/s) **(Hardware ?)**

Giganet - Throughput

- Transmission time in fiber optics cables

  tt = length of medium / (0,66 * c) with c = 300.000 km/s

  additionally delays in routers, switches etc.

  tt_{opt} = 100 km / (0,66 * 300.000 km/s) = 1/2000 s = 0,5 ms

  use path mtu discovery

  apply socket buffers to bandwidth delay product
- BDP = (B * RTT) = 622 Mb/s * 0.5 ms » 311 kb » 40 kB
- use setsockopt to set:
  - SO_SNDBUF und SO_RCVBUF 1 MB
  - TCP_NODELAY=1 and TCP_WINSHIFT=4

Giganet - Impediments

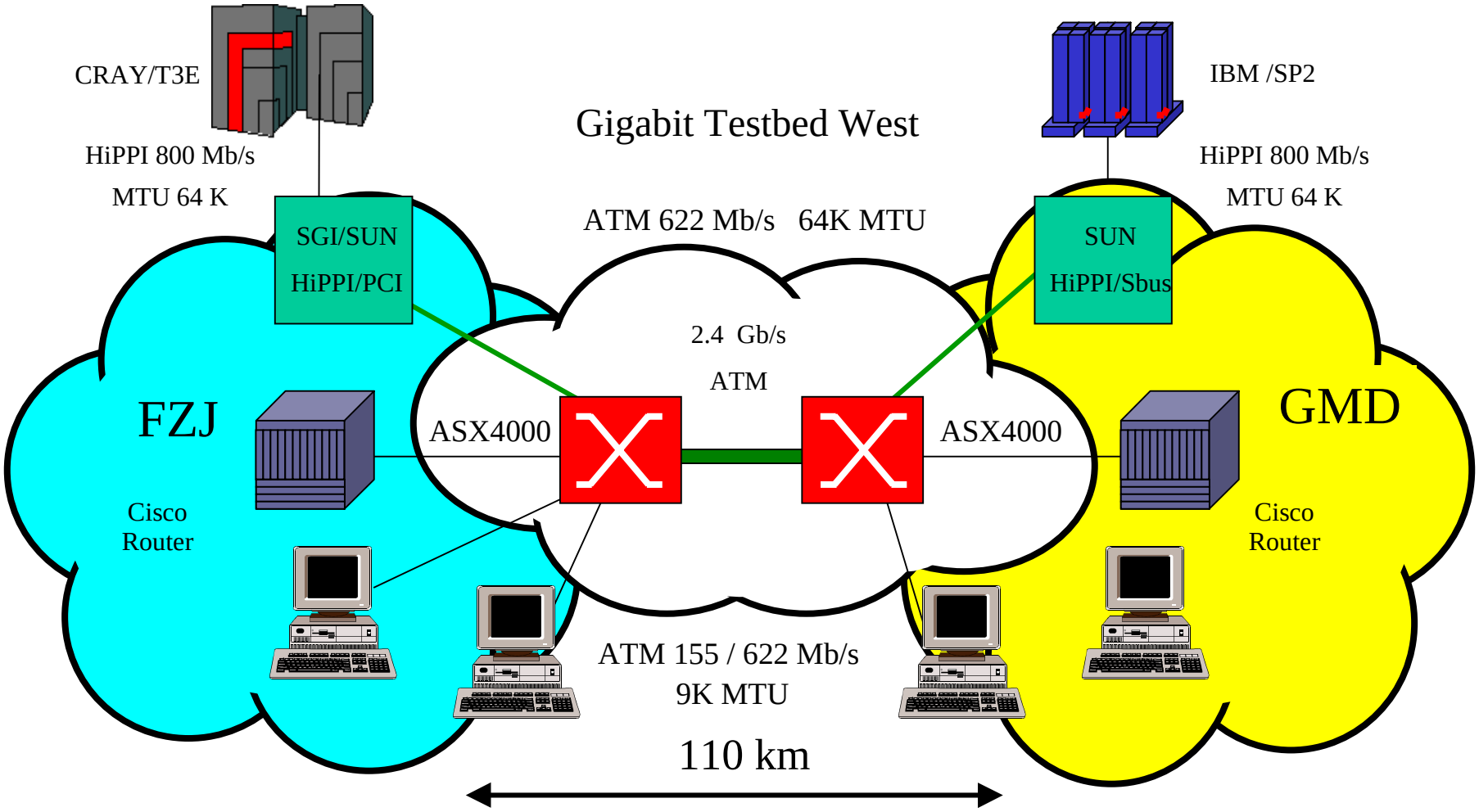
CRAY T3E communication throughput measured

- Maximum of 115 Mb/s via TCP/IP over ATM
  MTU 9180 (Default MTU from standard)
- Maximum of 430 Mb/s via TCP/IP over HiPPI
  MTU 64 KB because of IP-Header fields
- Maximum of 530 Mb/s via raw HiPPI
  no real MTU limitation

Netperf between SUN Ultra/60 and SGI Origin 200
maximum of 535 Mb/s user data via 622 Mb/s ATM

Gigabit Testbed West

Network Layout



Gigabit Testbed West

Connecting CRAY T3E and IBM SP2 via separate network

Problem:

- Interrupt rate of CRAY/T3E systems

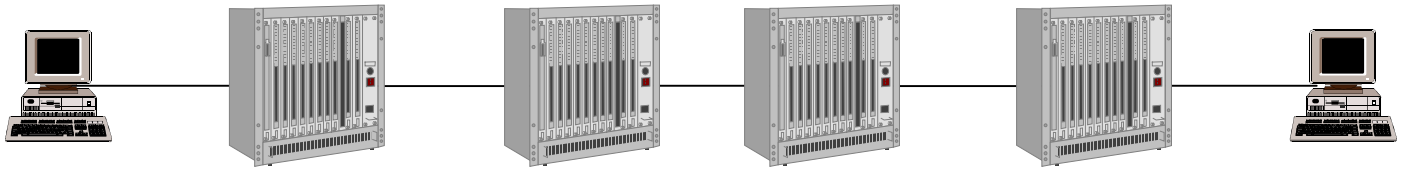
Solution:

Create two logical networks upon one physical network

- network 1 with 64k MTU between gateway systems (exact MTU 65280) as specified for CRAY systems on HiPPI networks
- network 2 with 9.180 MTU between directly connected ATM systems

Advantage:

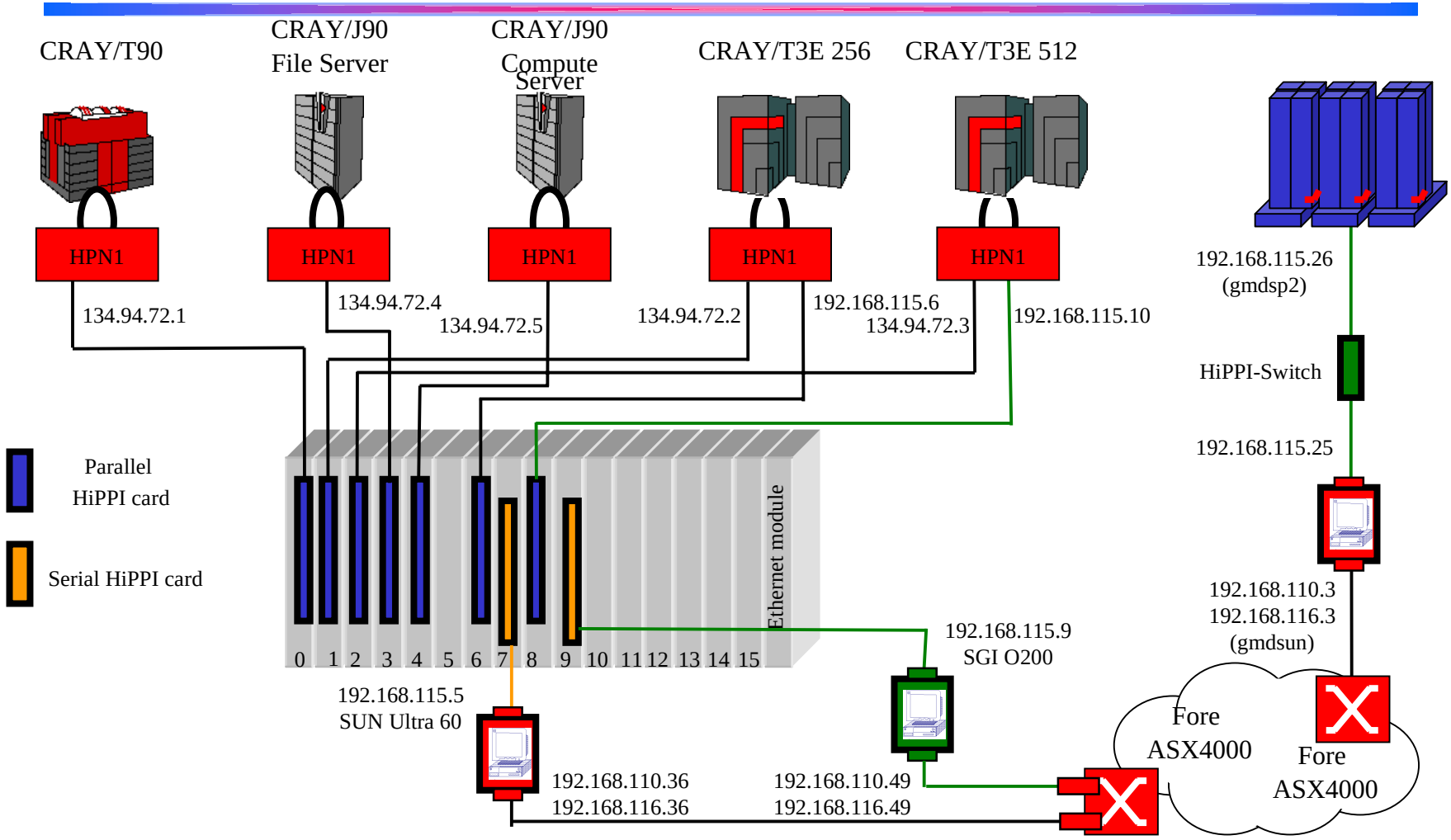
MTU-Path-Discovery on the end systems will find maximum value to use.



MTU: 9180 4356 1500 9180 65280

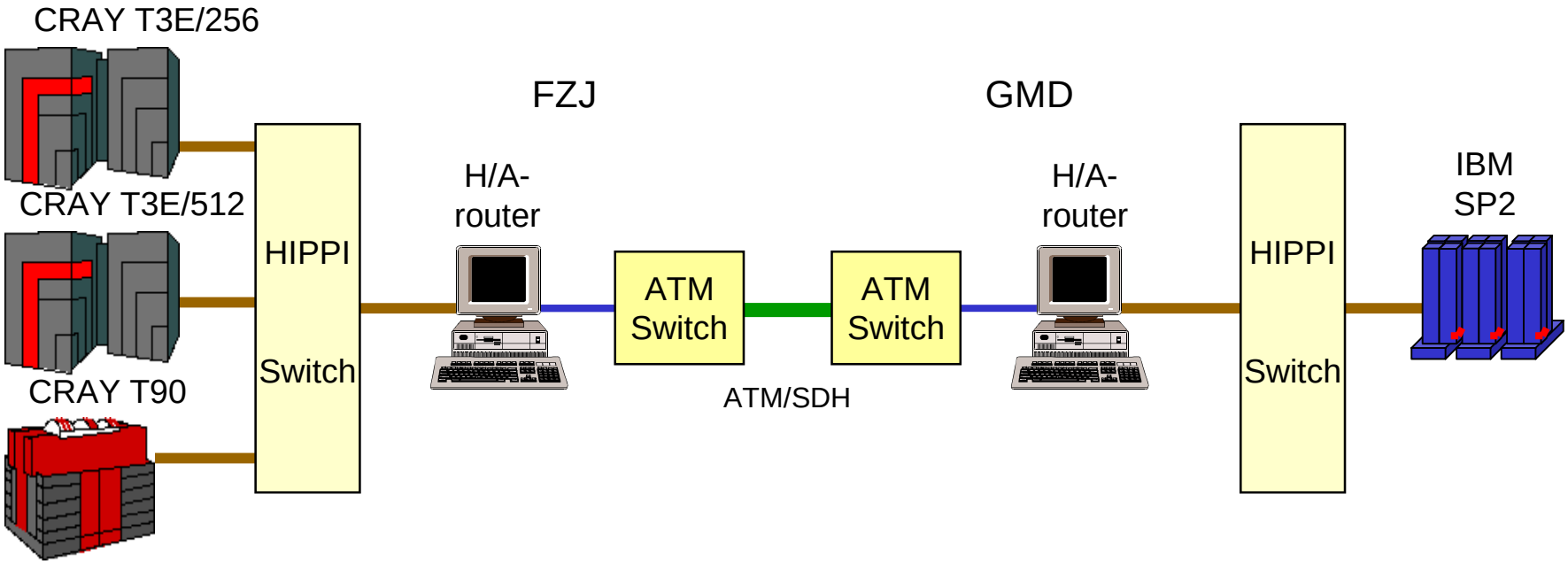
Status

CRAY HiPPI Testbed configuration



Communication nominal and real throughput

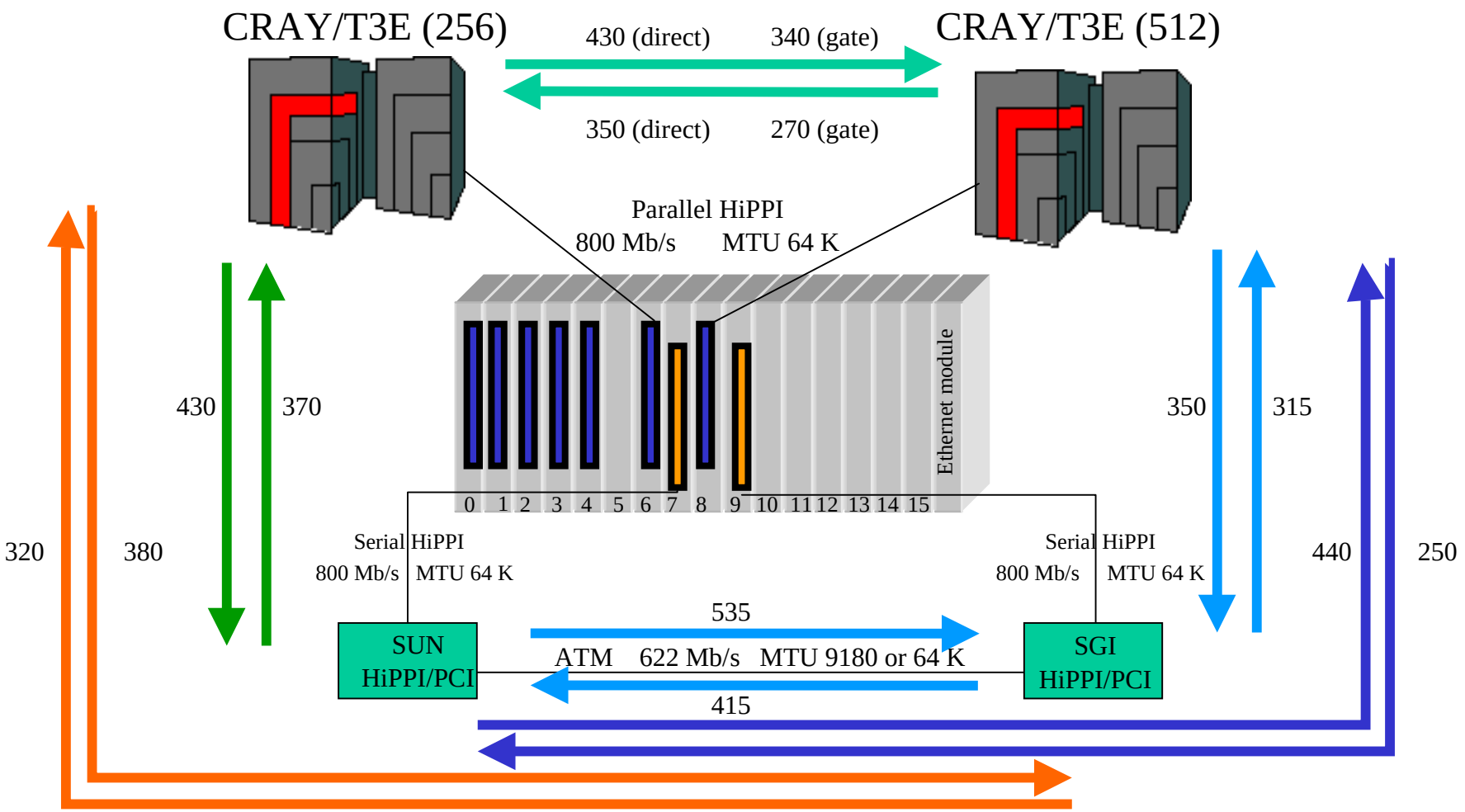
Nominal: 800 Mbps 800 Mbps 622 Mbps 2.4 Gbps 622 Mbps 800 Mbps 800 Mbps



Real: 430 Mbps 430 Mbps 530 Mbps 530 Mbps 530 Mbps 370 Mbps 370 Mbps

Gigabit Testbed West

TCP-Gateway-Layout (Beta-Tests in Jülich)



Future Tests

CRAY HiPPI Testbed configuration

- Solve HiPPI problem.
  Using large MTU sizes (65280 kB) does not work correctly
- Testing the other Cray Systems with HiPPI to ATM gateway (T90, J90)
- Testing different configurations if testbed is available
  - using 2 HPN1
  - using 2 Communication nodes within CRAY/T3E
  - using one Gateway for more than one machine
  - using same HiPPI device for local and remote communication
  - using multiple HiPPI devices for advanced throughput

Summary

- Time is ready for gigabit transmissions.
- Applications are capable using gigabit networks.
- Metacomputing may become reality in LAN as well as in WAN environments
- Therefore SGI/Cray has to prepare their systems with gigabit communication interfaces

„*The net is the computer and the computer is the net*“

((SuperComputer) Communications)
!= (Super (ComputerCommunications))