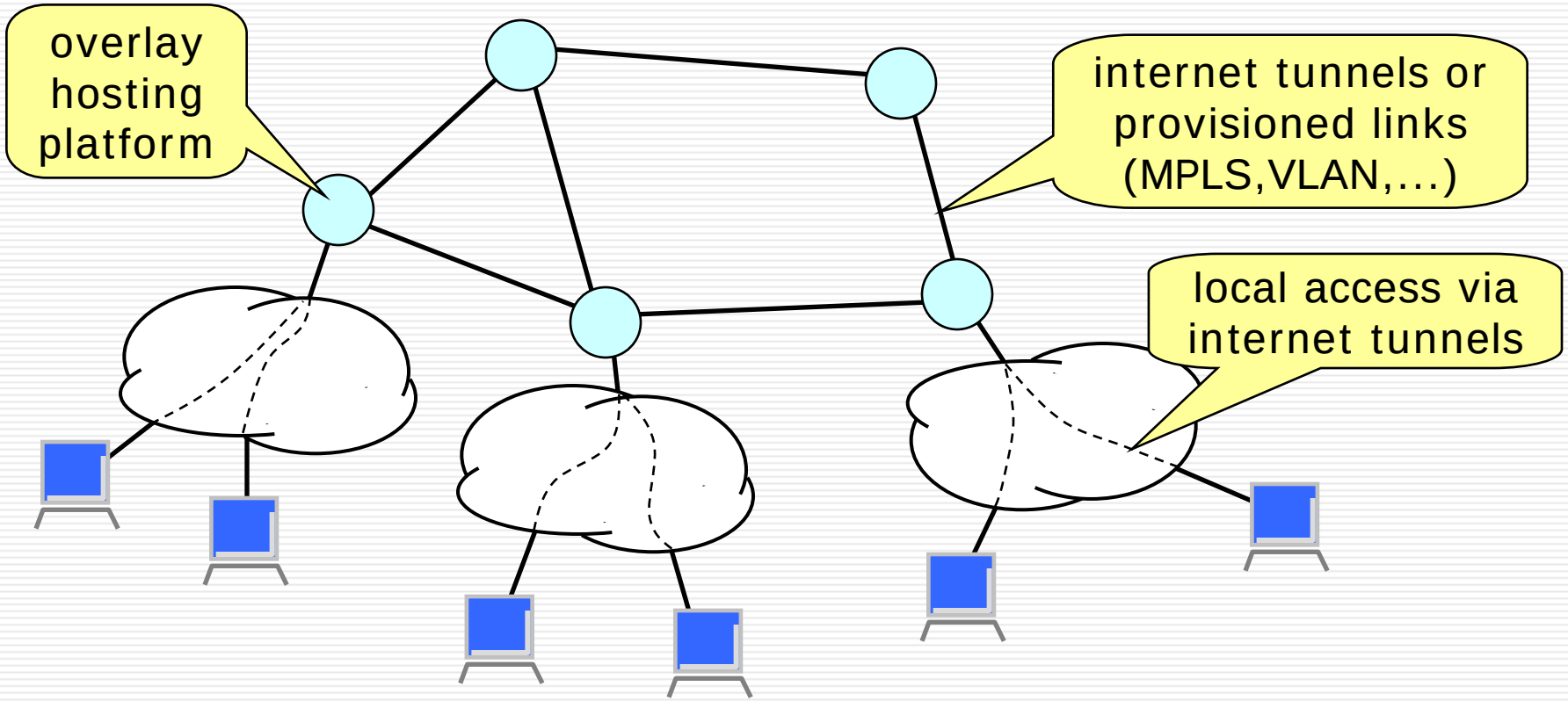
A Platform for High Performance Overlay Hosting Services

Jon Turner
with
Patrick Crowley, John DeHart, Brandon Heller,
Fred Kuhns, Sailesh Kumar, John Lockwood, Jing Lu,
Mike Wilson, Charlie Wiseman and Dave Zar

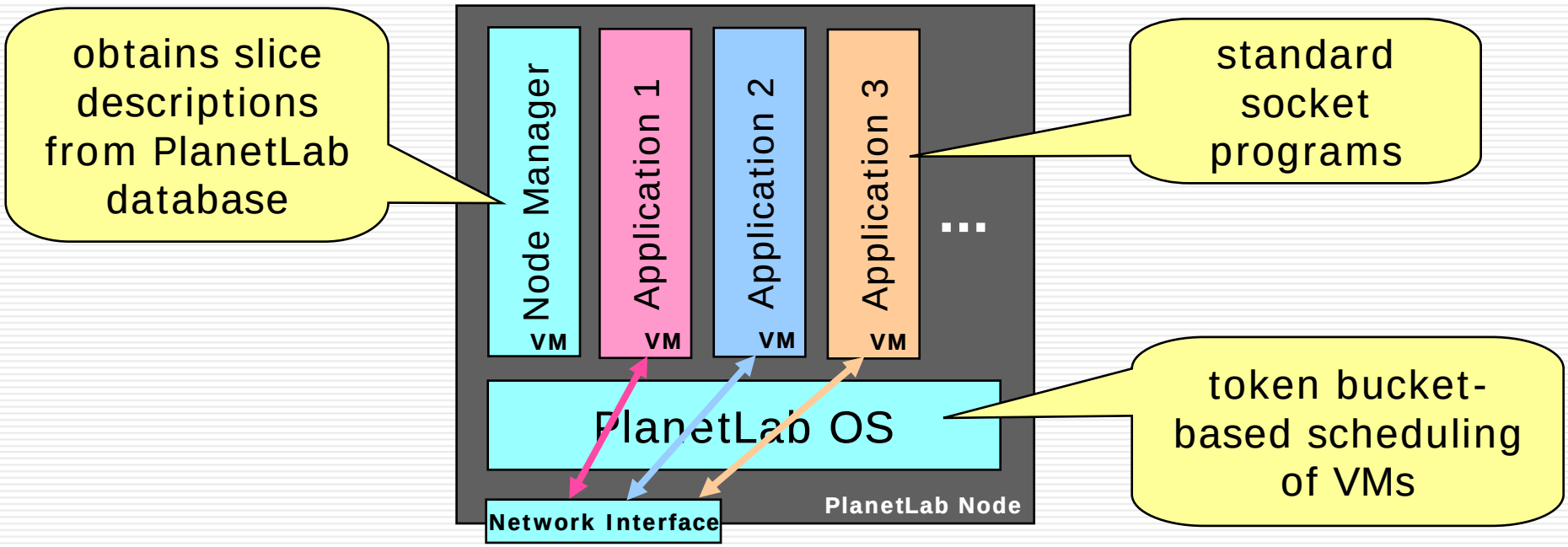
Overlay Hosting Services



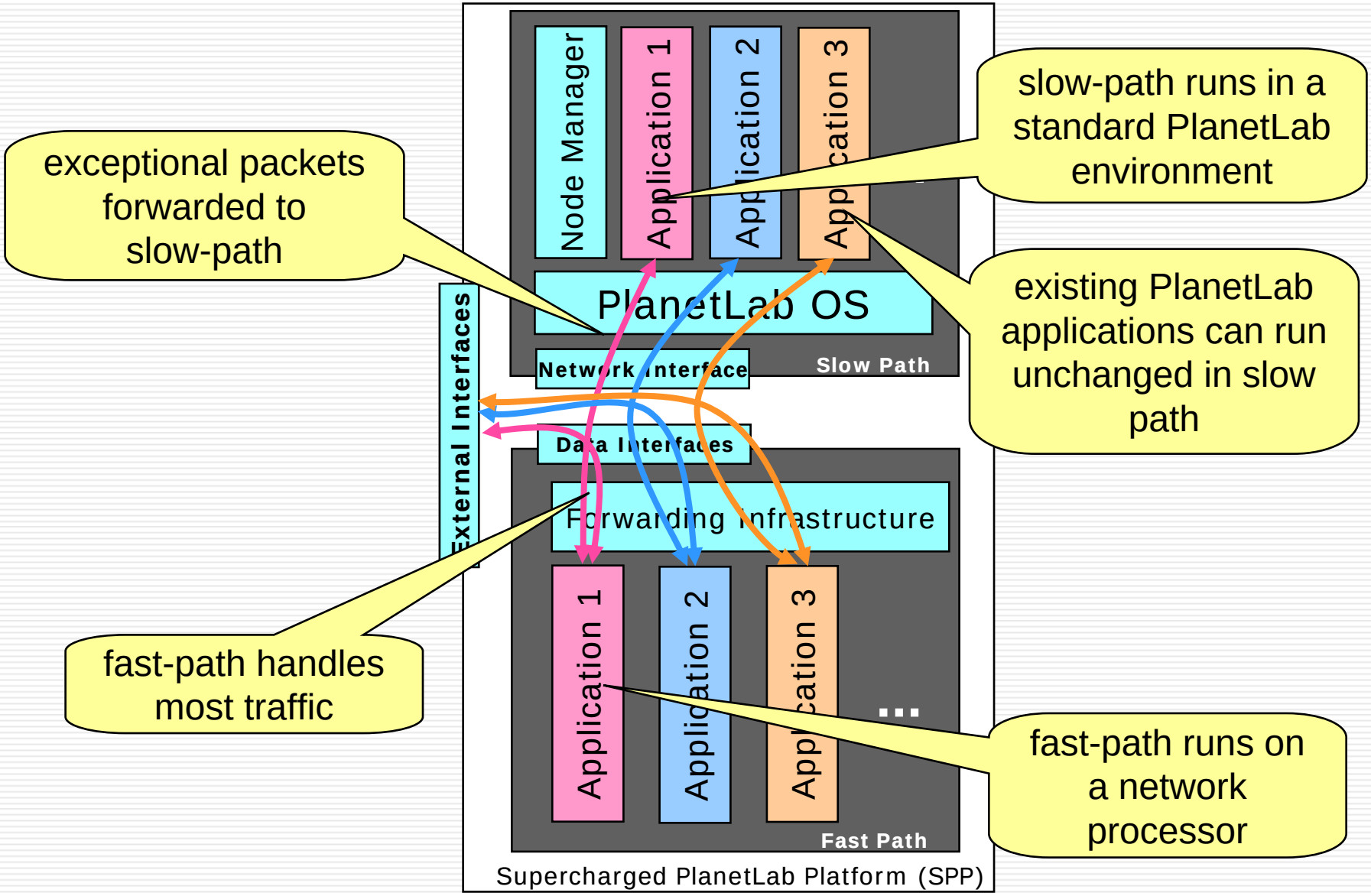
- Shared infrastructure hosting overlay-based services
- Objectives for overlay hosting platform
  - »internet-scale traffic volumes and consistently low latency
  - »deployment target: first Planetlab, then commercial and GENI

PlanetLab

- Canonical overlay hosting service, using PC platform
- Applications run as user-space processes in virtual machines
- Effective and important research testbed
- But, low throughput and widely variable latency limits its potential as service deployment platform



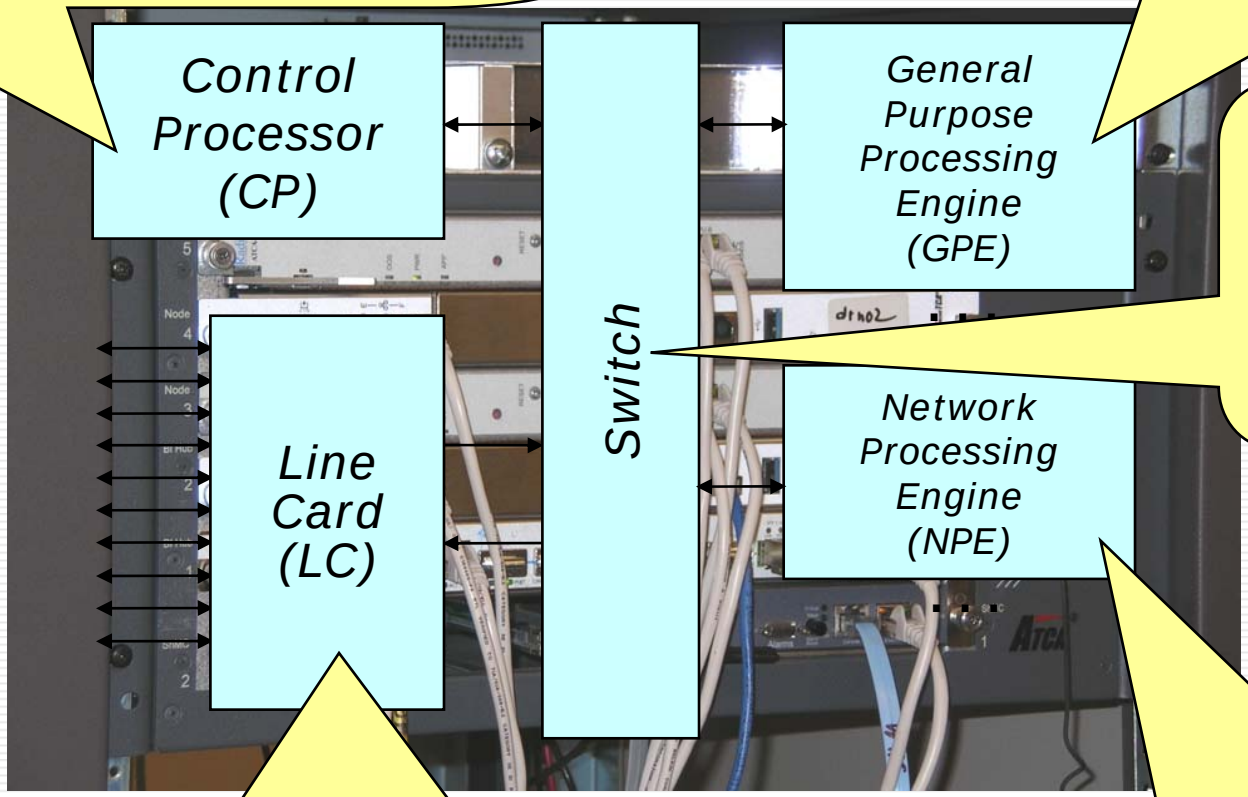
Supercharging PlanetLab



SPP Components

conventional server which coordinates system components and synchronizes with PlanetLab

conventional server blades supporting standard PlanetLab environment

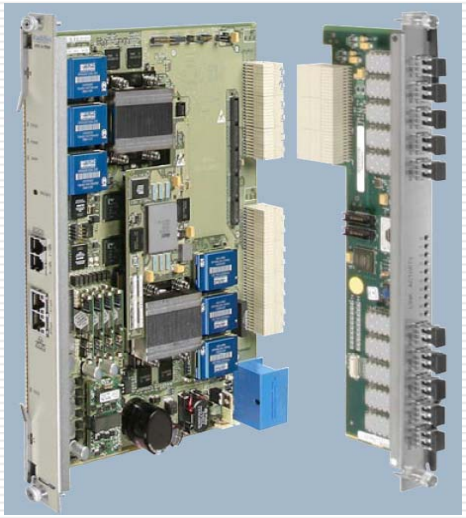


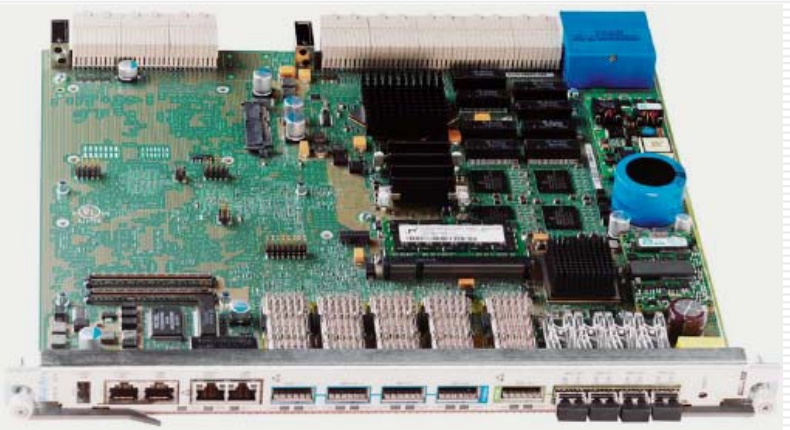
blade containing 10GE data switch and 1GE control switch

dual Intel IXP 2850 blade which forwards packets to correct PEs

dual Intel IXP 2850 blades supporting application fast-paths

ATCA Boards



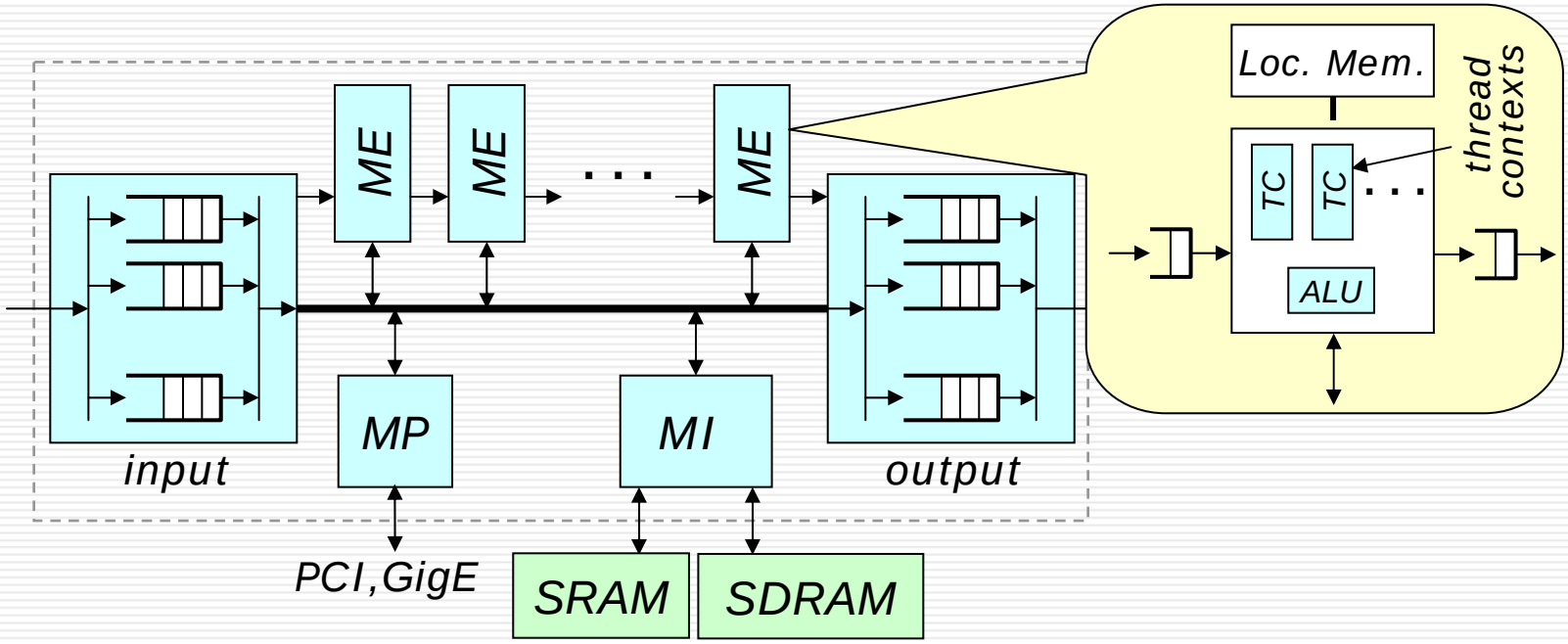


- Intel server blades
  - » for CP and GPE
  - » dual Xeons (2 GHz)
  - » 4x1GbE
  - » on-board disk
  - » Advanced Mezzanine Card slot
- Radisys NP blades
  - » for LC and NPE
  - » dual IXP 2850 NPs
    - 3xRDRAM
    - 4xSRAM
    - shared TCAM
  - » 2x10GbE to backplane
  - » 10x1GbE external IO (or 1x10GbE)
- Radisys switch blade
  - » up to 16 slot chassis
  - » 10 GbE fabric switch
  - » 1 GbE control switch
  - » full VLAN support
- Scaling up
  - » 5x10 GbE to front
  - » 2 more to back

What You Need to Build Your Own

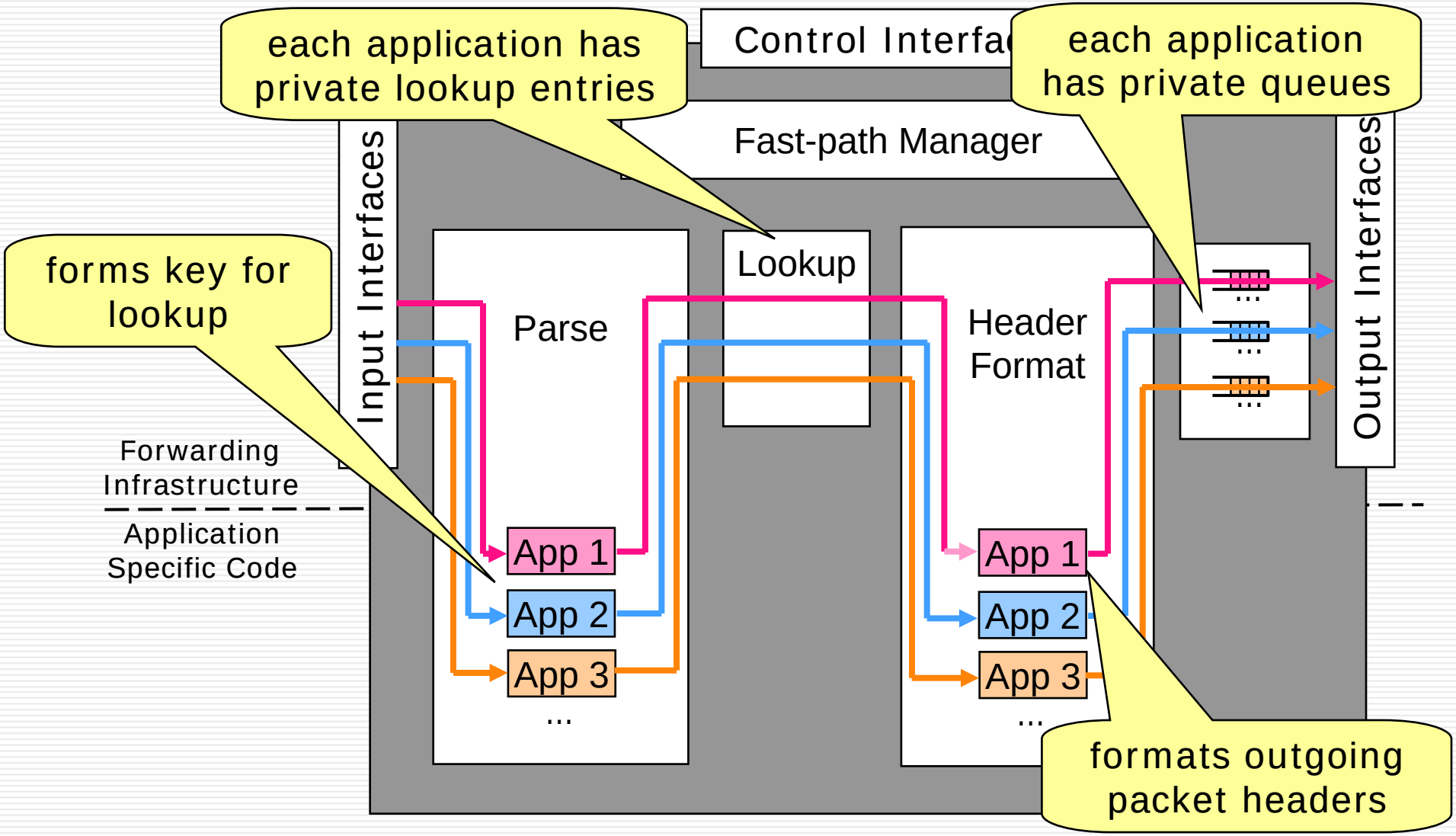
Qty	Description	Supplier	Model
1	Dual Network Processor Module with IO	Radisys	A7K-PPM10-CFG002
2	Dual Network Processor Module		A7010-BASE-2855
2	18 MB IDT TCAM Module		A7010-TCAM-01-R
3	10 Gb/s Fabric Interface Card		A7010-FIC-2X10G
1	10 GE/1GE Switch & Control Module		A2210-SWH-CFG-01
1	RTM with extra IO ports		A5010-SPM-01
5	1GE plugin modules (4 per kit)		A2K-SFP-C
2	Server blade with 2 dual-core Xeon processors	Intel	MPCBL004N01Q
1	Zephyr 6 Slot ATCA Shelf	Schroff	ZR5ATC6TMDPEM2N
1	Shelf Manager		21593-375
1	Alarm Board		ISAP2
1	1U Power Supply Shelf	Unipower	TPCPR1U3B
1	48 Vdc/25A Power Supply		TPCP7000
1	115 Vac/15A Power Cord		364-1409-0000

IXP 2850 Overview

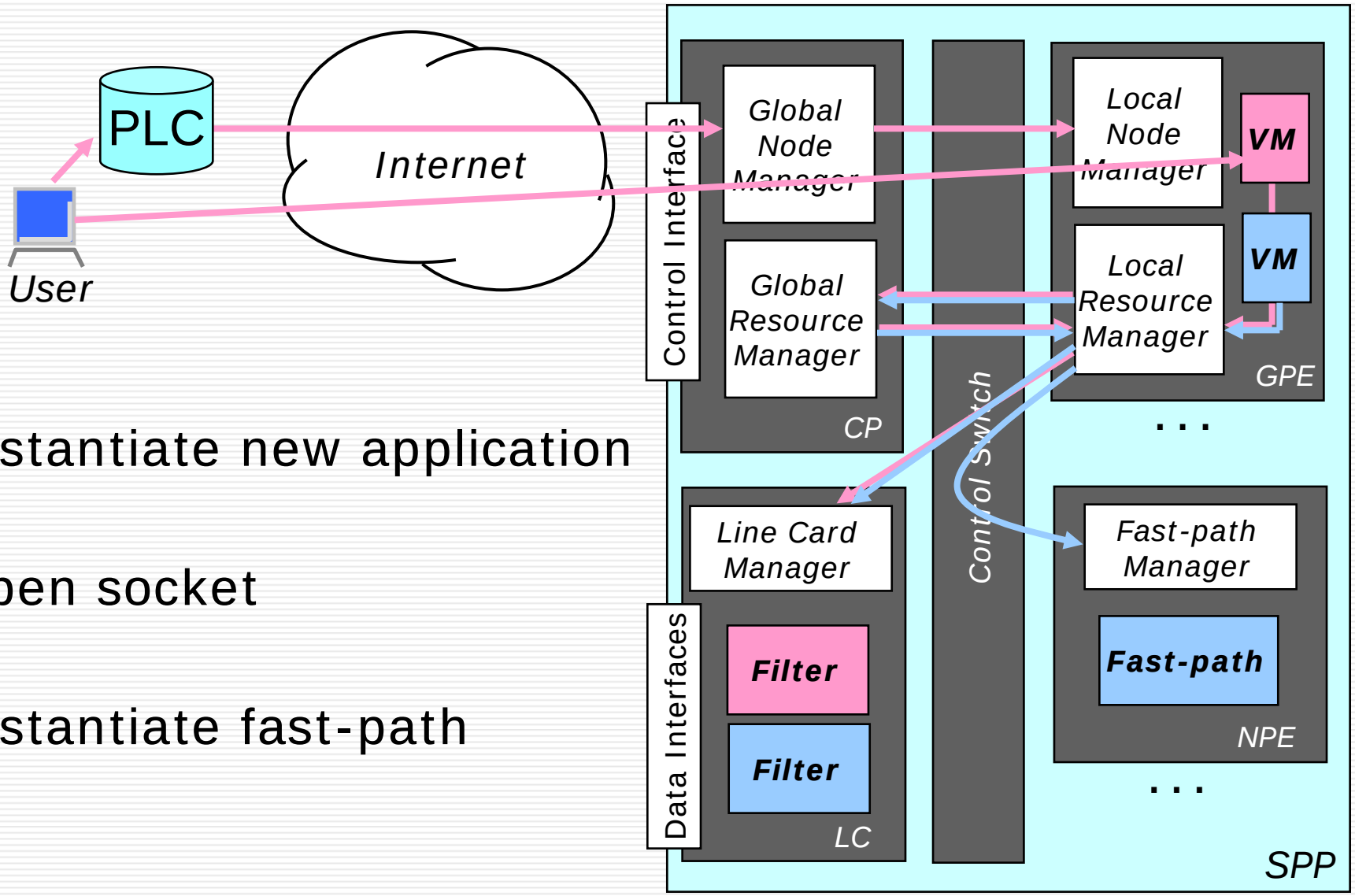


- 16 multi-threaded MicroEngines (MEs)
  - » 8 thread contexts with rapid switching capability
  - » fast nearest-neighbor connections for pipelined apps
- 3 SDRAM and 4 SRAM channels (optional TCAM)
- Management Processor (MP) for control

Sharing the NPE

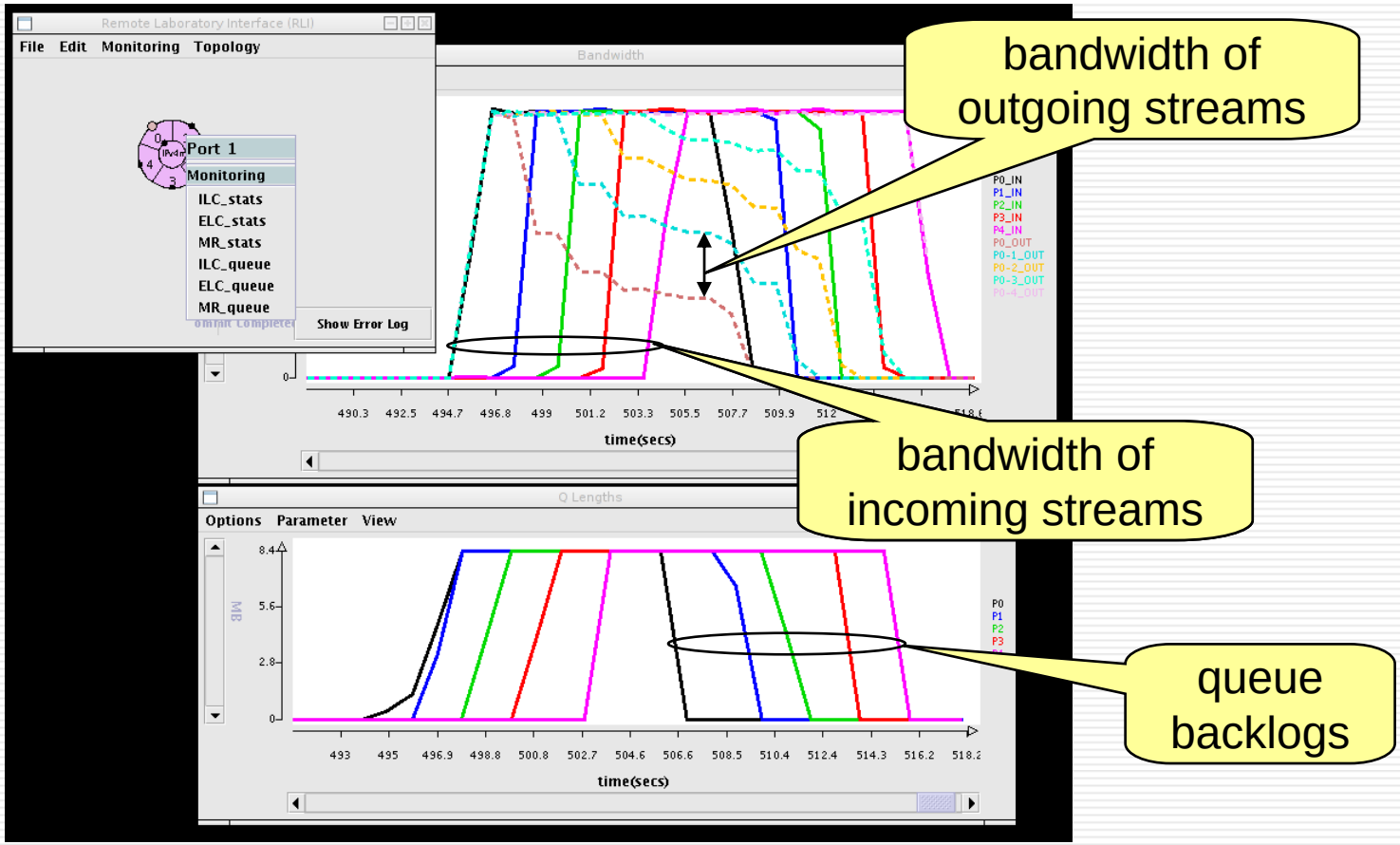


System Control



- Instantiate new application
- Open socket
- Instantiate fast-path

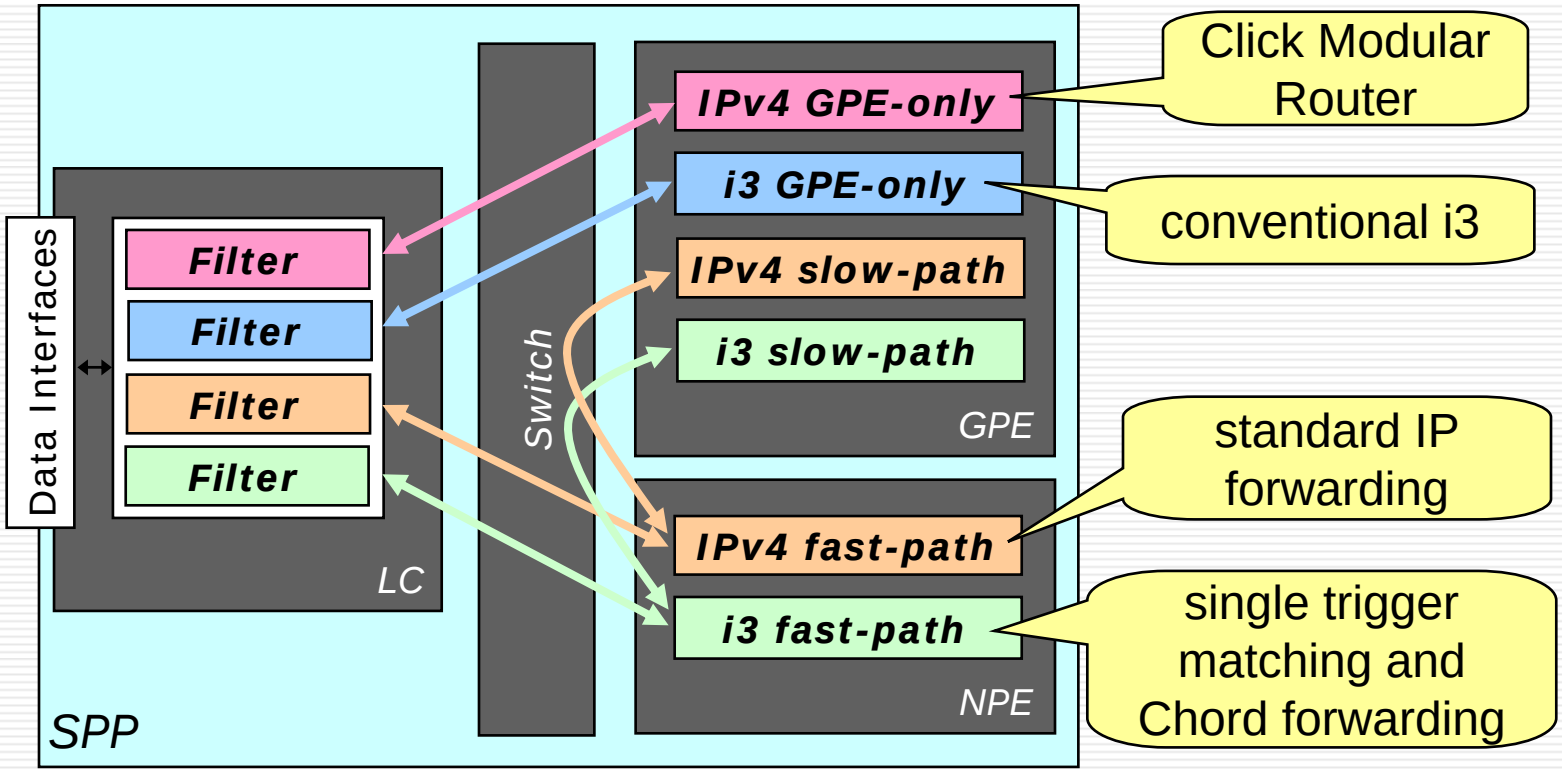
Basic Operational Demo



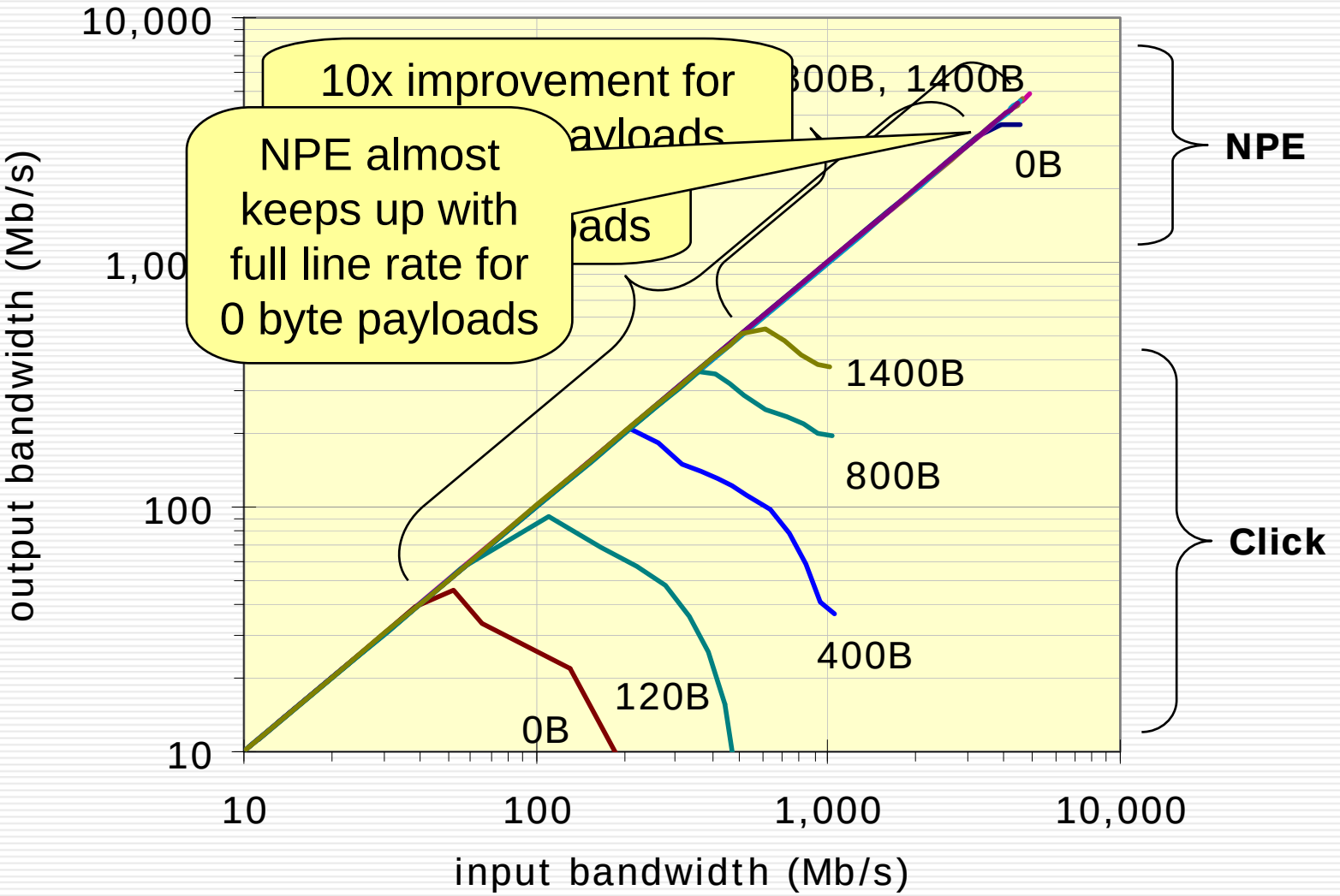
- Traffic on 5 inputs going to single output
  - » offset, but overlapping traffic bursts
- Each flow has different share of output bandwidth

Evaluation

- Slice 1 – IPv4
  - » packets arrive/depart in UDP tunnels
- Slice 2 – Internet Indirection Infrastructure (i3)
  - » packets contain *triggers* matched to IP addresses
  - » no match at local node results in Chord forwarding



IPv4 Throughput Comparison

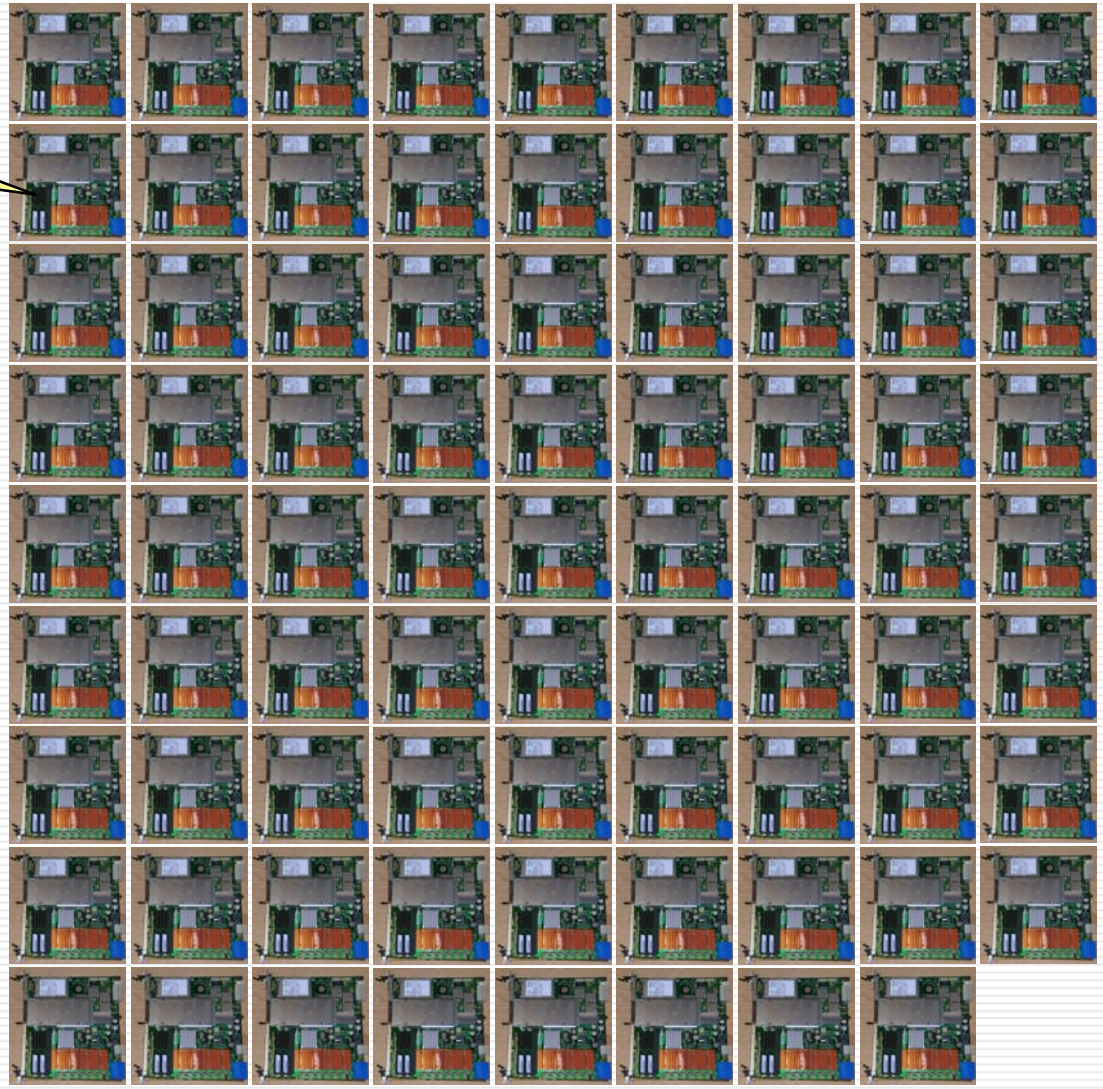


So, what this means is...

Xeon Server blade

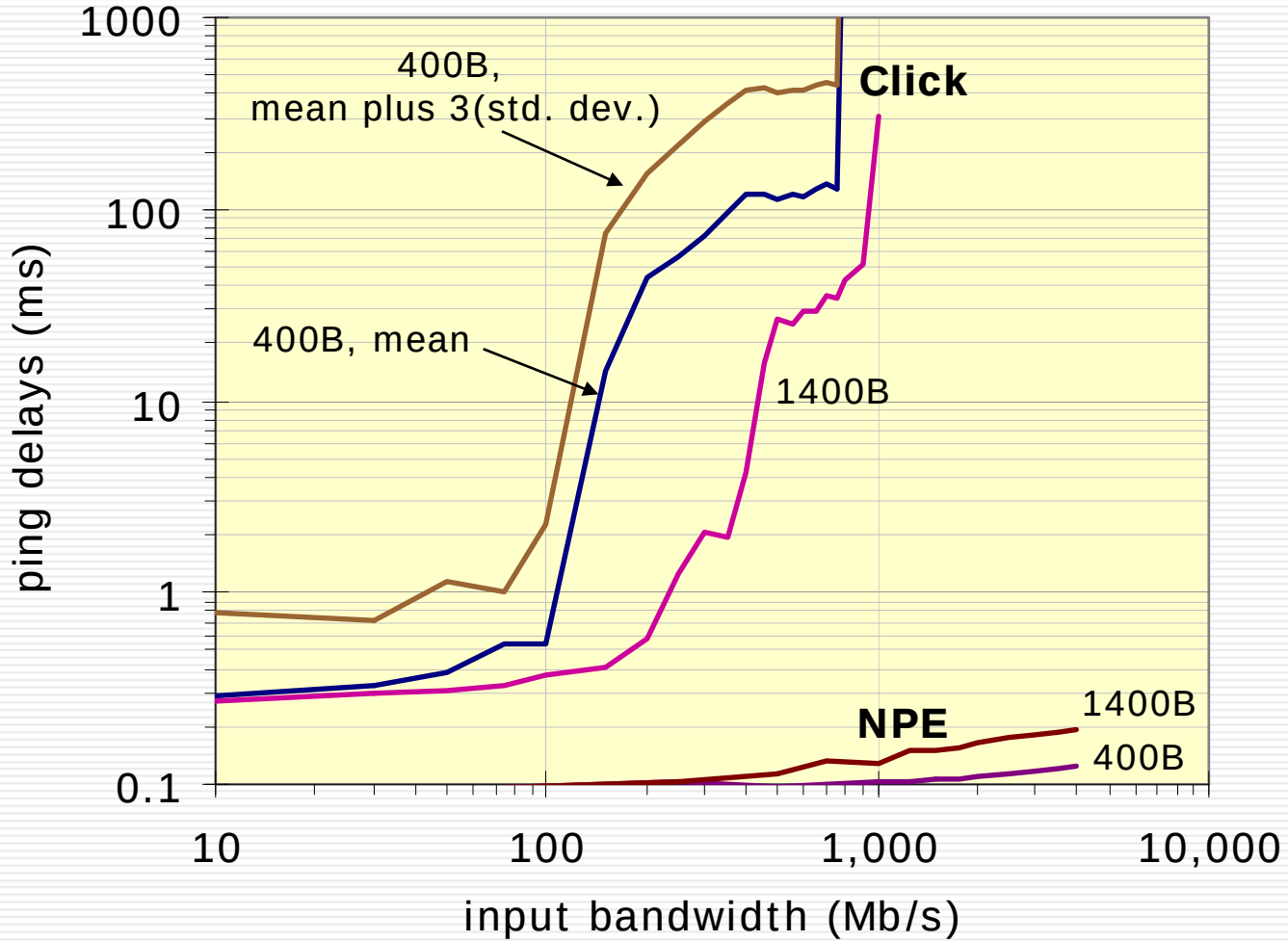
IXP 2850 NP blade

=



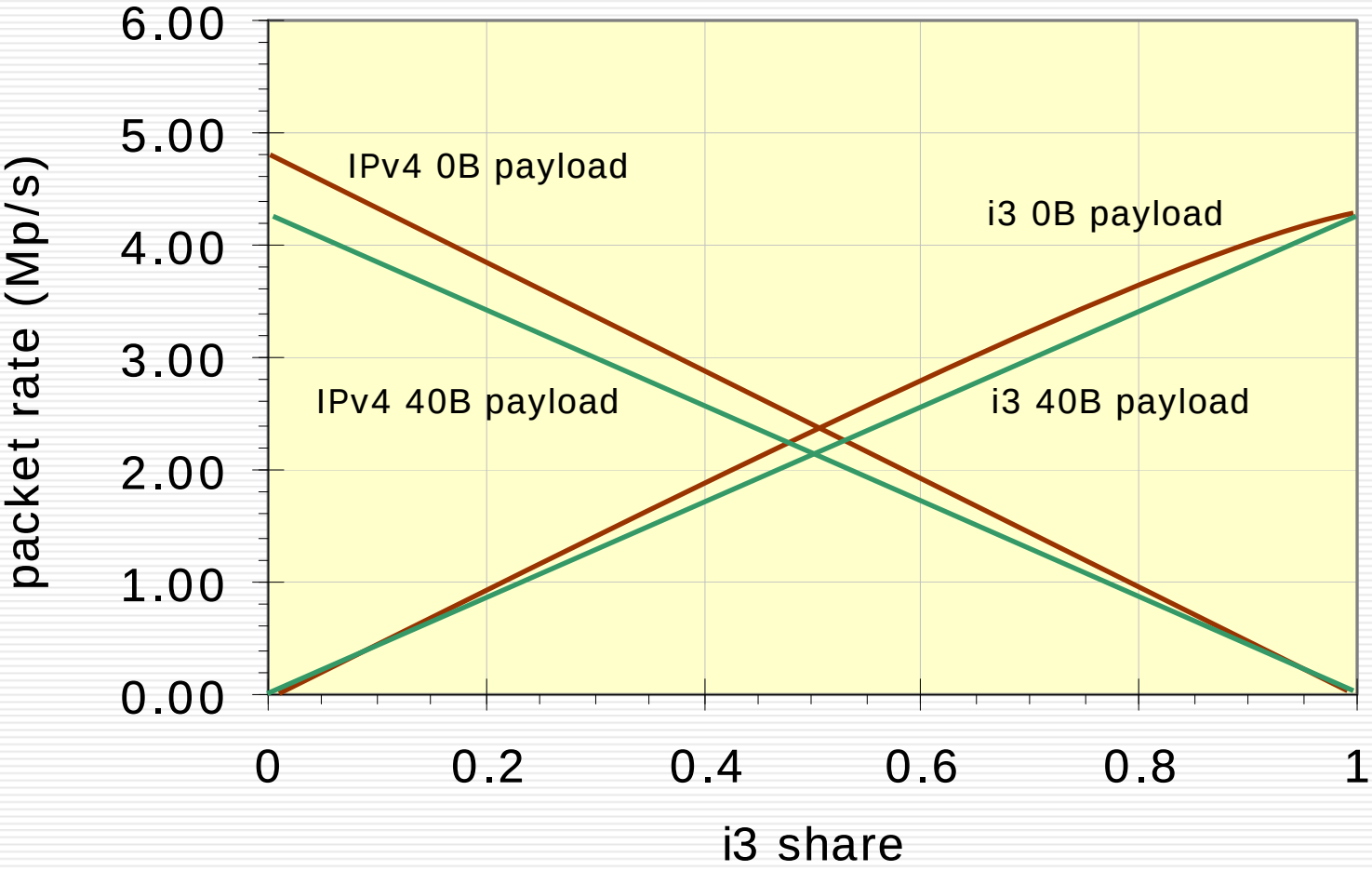
price-performance advantage of >15X

IPv4 Latency Comparison



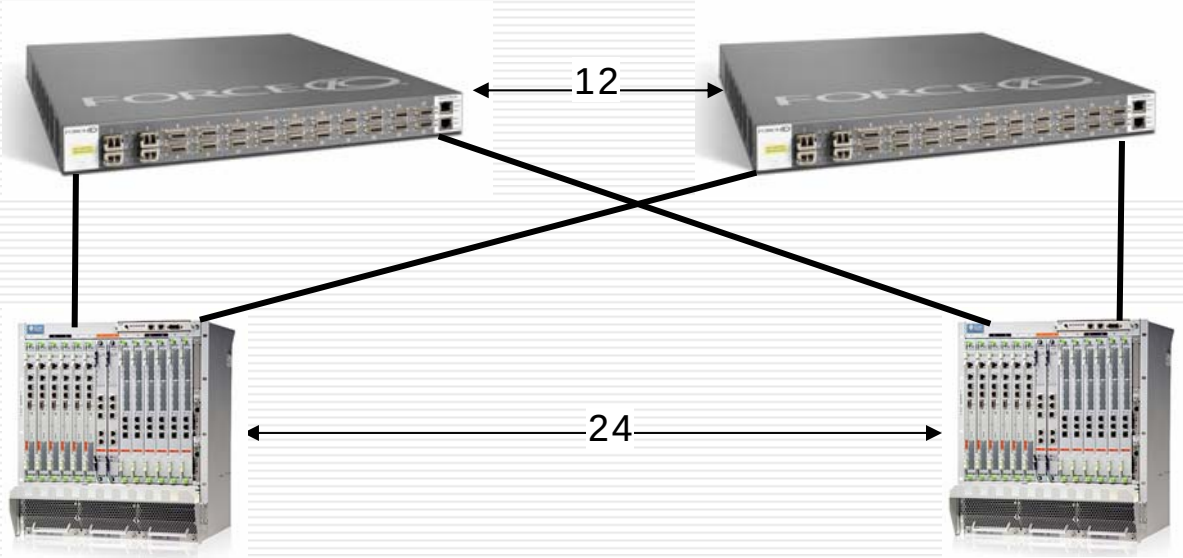
- 8 IPv4 instances
- Measured ping delay against background traffic

IPv4/i3 Fast-Path Throughput Comparison



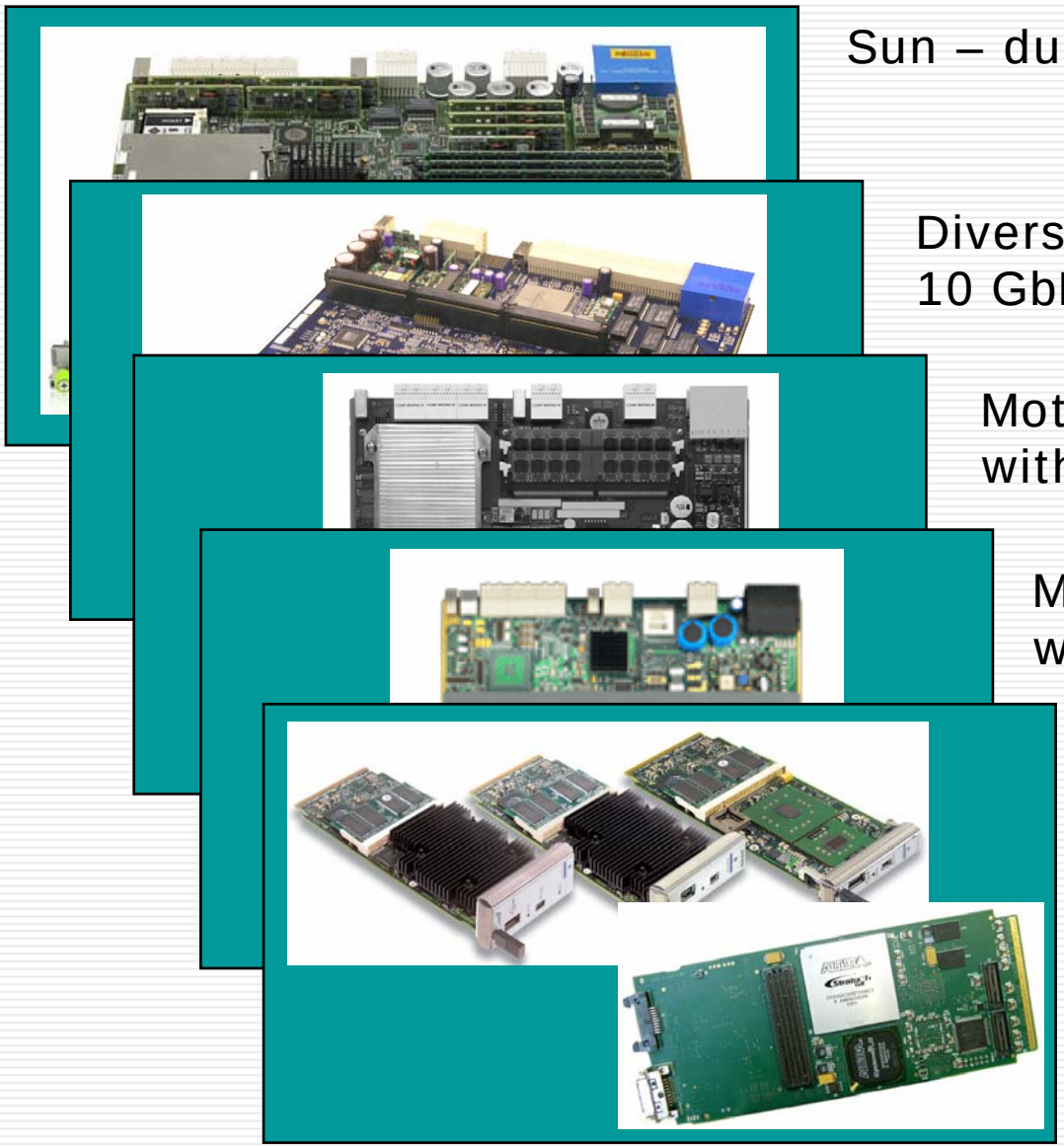
- Constant input rate of 5 Gb/s

Scaling Up



- 14 slot chassis
  - »3 Line Cards
  - »2 switch blades
  - »9 processing blades (NP or server)
- Multi-chassis systems
  - »direct connection using expansion ports – up to 7 chasses
  - »indirect connection using separate 10 GbE switches – up to 24 chasses

Other ATCA Components



Sun – dual 8-core UltraSparc

Diversified Technologies
10 GbE Switch Blade

Motorola – dual Cavium NPs
with 16 MIPS cores each

Motorola – 4xAMC carrier,
with 10Gx1G switch

Artesyn - AMC Compute blades

Bitstream - AMC FPGA blades

Summary

- ATCA is important enabling development for overlay hosting services like GENI
  - » market for open, programmable network subsystems
  - » many vendors, variety of products
  - » greater opportunity for network service innovation
- Growing role of multi-core processors
  - » to use them effectively, must design for parallelism
  - » requires deeper understanding of performance
- Conventional servers have dreadful performance on IO-intensive applications
  - » partly hardware, but mostly software
  - » to fix, need to push fast-path down into drivers and program for multi-core parallelism