PetaCache: Data Access Unleashed

Tofigh Azemoon, Jacek Becla, Chuck Boeheim,
Andy Hanushevsky, David Leith, Randy Melen,
Richard P. Mount, Teela Pulliam, William Weeks

Stanford Linear Accelerator Center

September 3, 2007

Outline

- Motivation – Is there a Problem?
- Economics of Solutions
- Practical Steps – Hardware/Software
- Some Performance Measurements

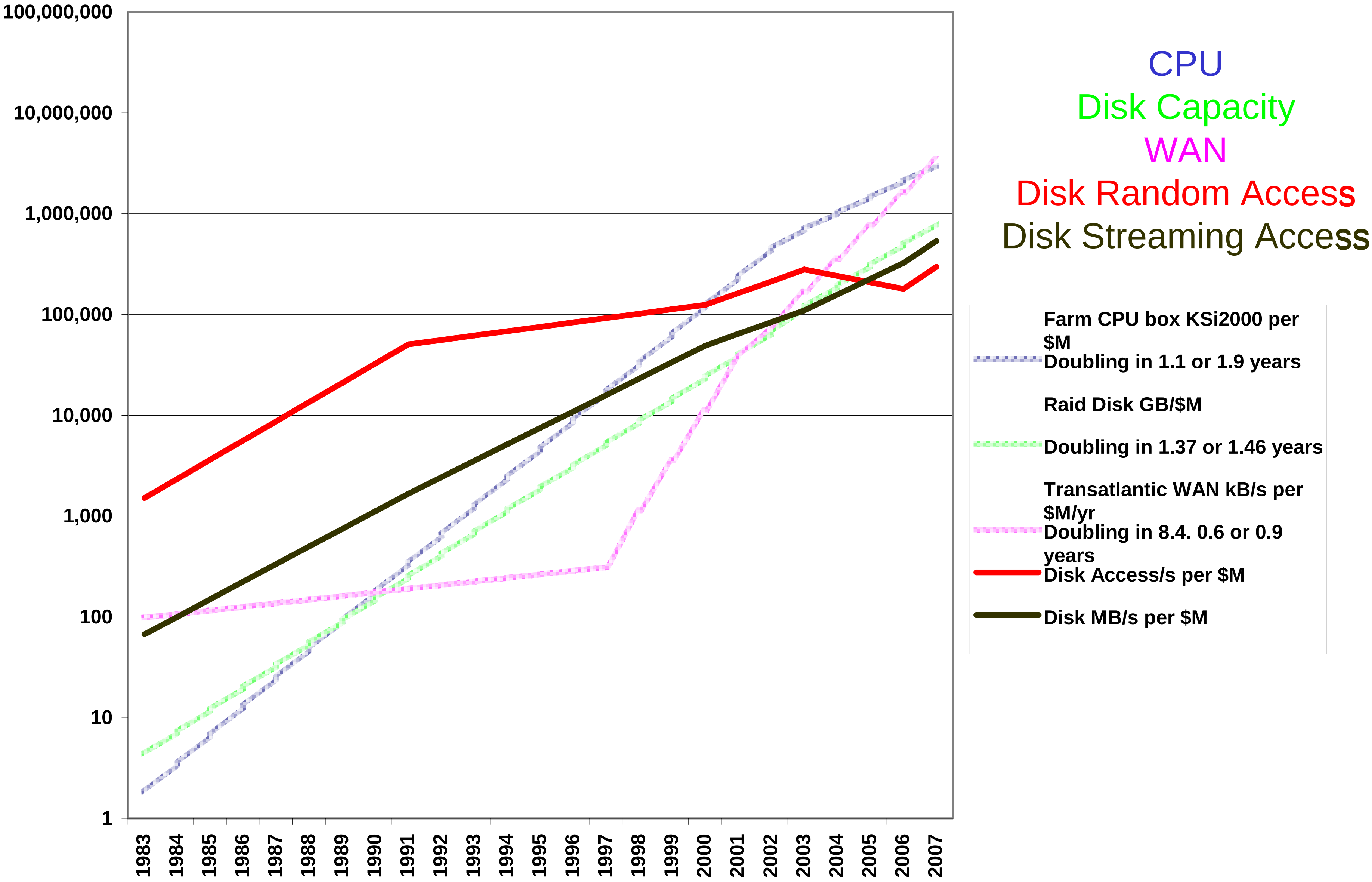
Motivation

Storage In Research:
Financial and Technical Observations

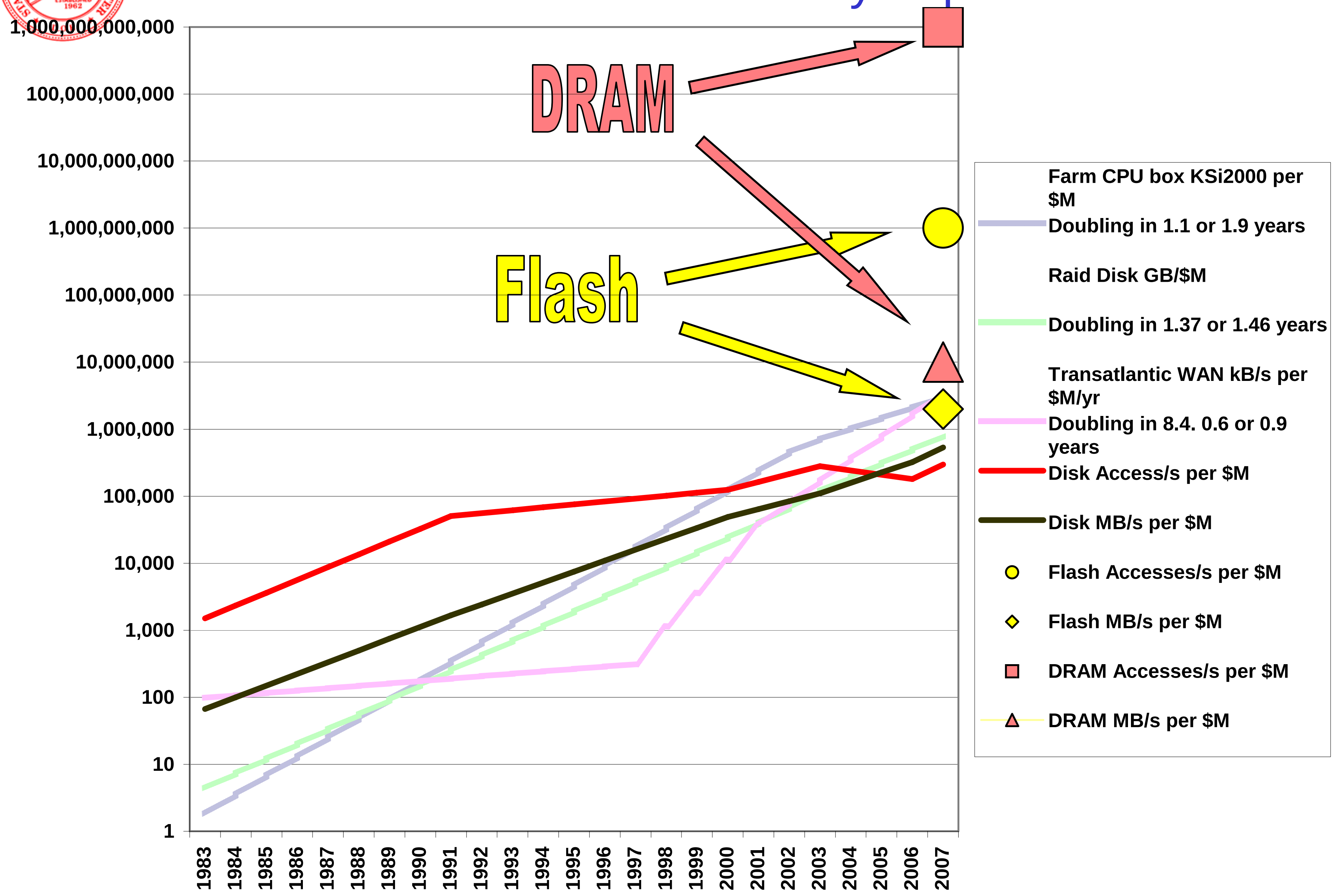
- Storage costs often dominate in research
  - CPU per $ has fallen faster than disk space per $ for most of the last 25 years
- Accessing data on disks is increasingly difficult
  - Transfer rates and access times (per $) are improving more slowly than CPU capacity, storage capacity or network capacity.
- The following slides are based on equipment and services that I* have bought for data-intensive science

* The WAN services from 1998 onwards were bought by Harvey Newman of Caltech

Price/Performance Evolution: My Experience



Price/Performance Evolution: My Experience



Another View

- In 1997 $M bought me:

  ~ 200-core CPU farm
  (~few x 10^8 ops/sec/core)

  or

  ~ 1000-disk storage system
  (~2 x 10^3 ops/sec/disk)

- Today $1M buys me (you):

  ~ 2500-core CPU farm
  (~few x 10^9 ops/sec/core)

  or

  ~ 2500-disk storage system
  (~2 x 10^3 ops/sec/disk)

- In 5 – 10 years ?

Impact on Science

- Sparse or random access must be derandomized
- Define, in advance, the interesting subsets of the data
- Filter (skim, stream) the data to instantiate interest-rich subsets

Economics of Solutions

Economics of LHC Computing

- Difficult to get $10M additional funding to improve analysis productivity
- Easy to re-purpose $10M of computing funds if it would improve analysis productivity

Cost-Effectiveness

- DRAM Memory:
  – $100/gigabyte
  – SLAC spends ~12% of its hardware budget on DRAM
- Disks (including servers)
  – $1/gigabyte
  – SLAC spends about 40% of its hardware budget on disk
- Flash-based storage (SLAC design)
  – $10/gigabyte
  – If SLAC had been spending 20% of its hardware budget on Flash we would have over 100TB today.

Practical Steps

The PetaCache Project

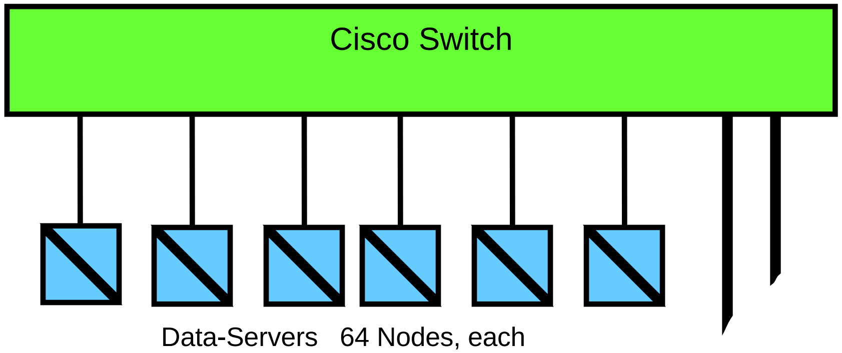
PetaCache Goals

1. Demonstrate a revolutionary but cost effective new architecture for science data analysis
2. Build and operate a machine that will be well matched to the challenges of SLAC/Stanford science

The PetaCache Story So Far

- We (BaBar, HEP) had data-access problems
- We thought and investigated
  - Underlying technical issues
  - Broader data-access problems in science
- We devised a hardware solution
  - We built a DRAM-based prototype
  - We validated the efficiency and scalability of our low-level data-access software, xrootd
  - We set up a collaboration with SLAC's electronics wizards (Mike Huffer and Gunther Haller) to develop a more cost-effective Flash-based prototype
- We saw early on that new strategies and software for data access would also be needed

DRAM-Based Prototype Machine (Operational early 2005)



Data-Servers 64 Nodes, each
Sun V20z, 2 Opteron CPU, 16 GB memory
1TB total Memory
Solaris or Linux (mix and match)

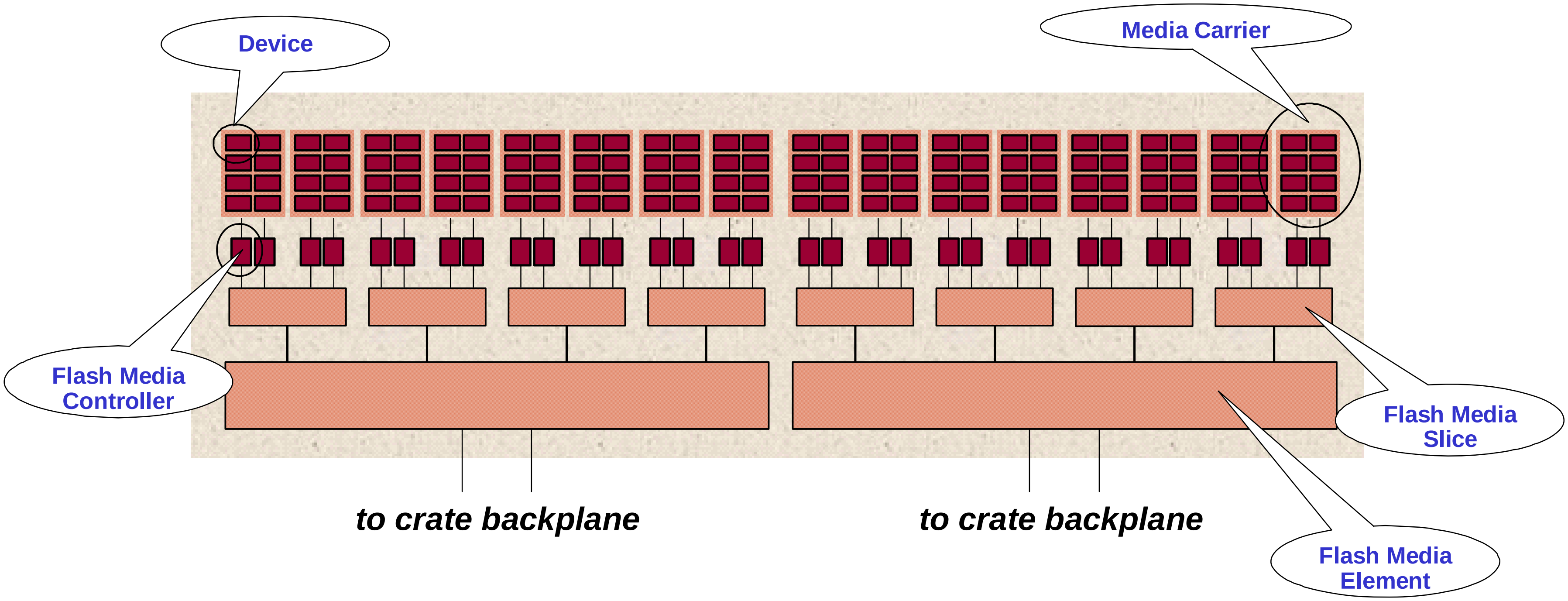
PetaCache
MICS + HEP-
BaBar Funding

DRAM-Based Prototype

FLASH-Based Prototype Operational Real Soon Now

- 5 TB of Flash memory
- Fine-grained, high bandwidth access

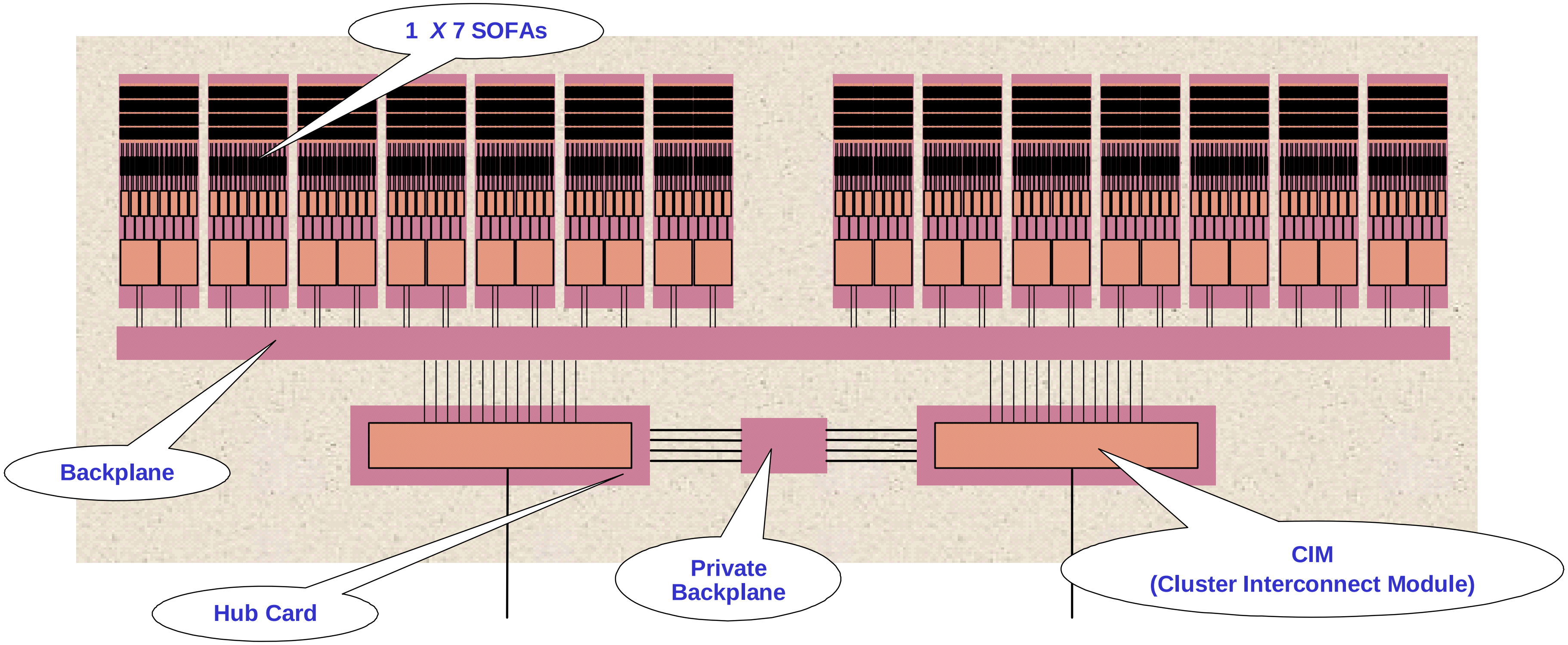
Building blocks (the SOFA)



Block	# of devices	GBytes
Device	*1*	*4*
Media *Carrier*	*8*	*32*
Flash Media *Slice*	*16*	*64*
Flash Media *Element*	*64*	*256*
SOFA	*128*	*512*

Slide from Mike Huffer

Building blocks (the Crate)

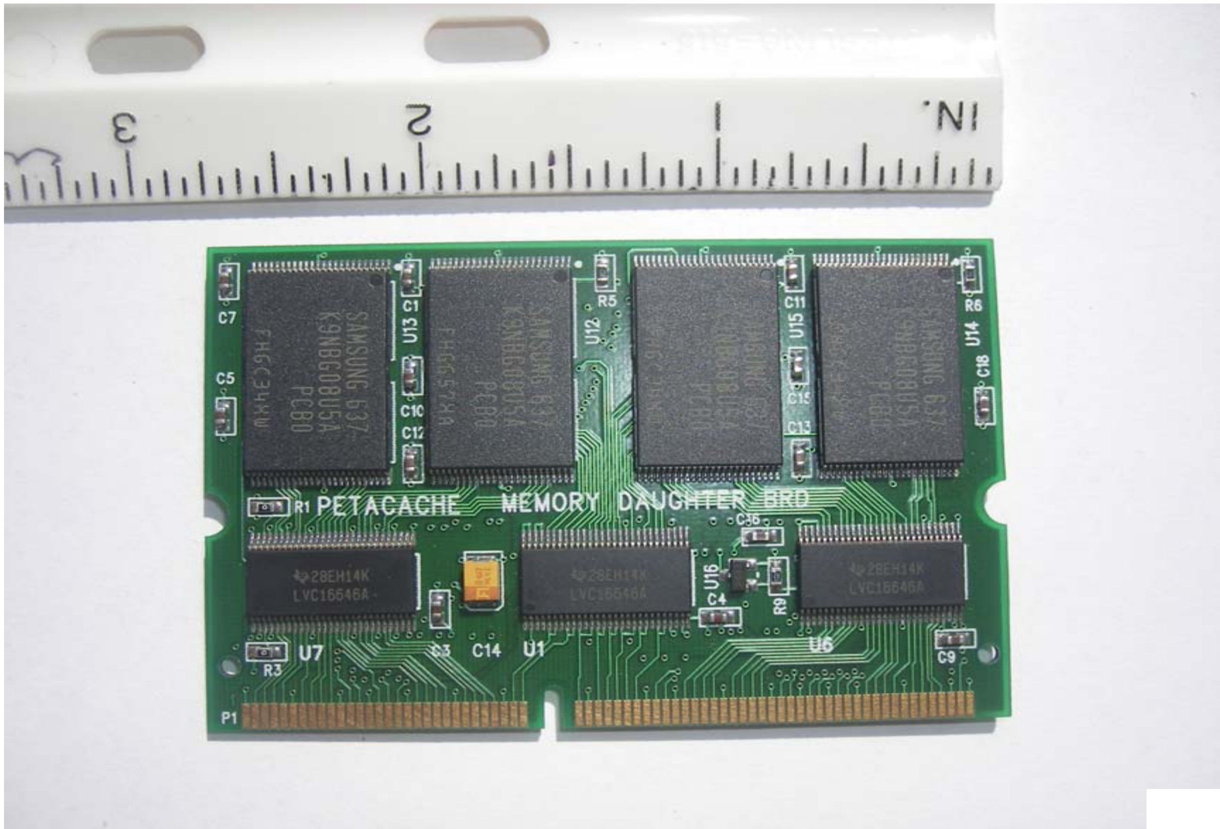


Block	# of devices	TBytes
SOFA	*128*	*0.5*
Hub Card	*896*	*3.5*
Crate	*1792*	*7.0*

Slide from Mike Huffer

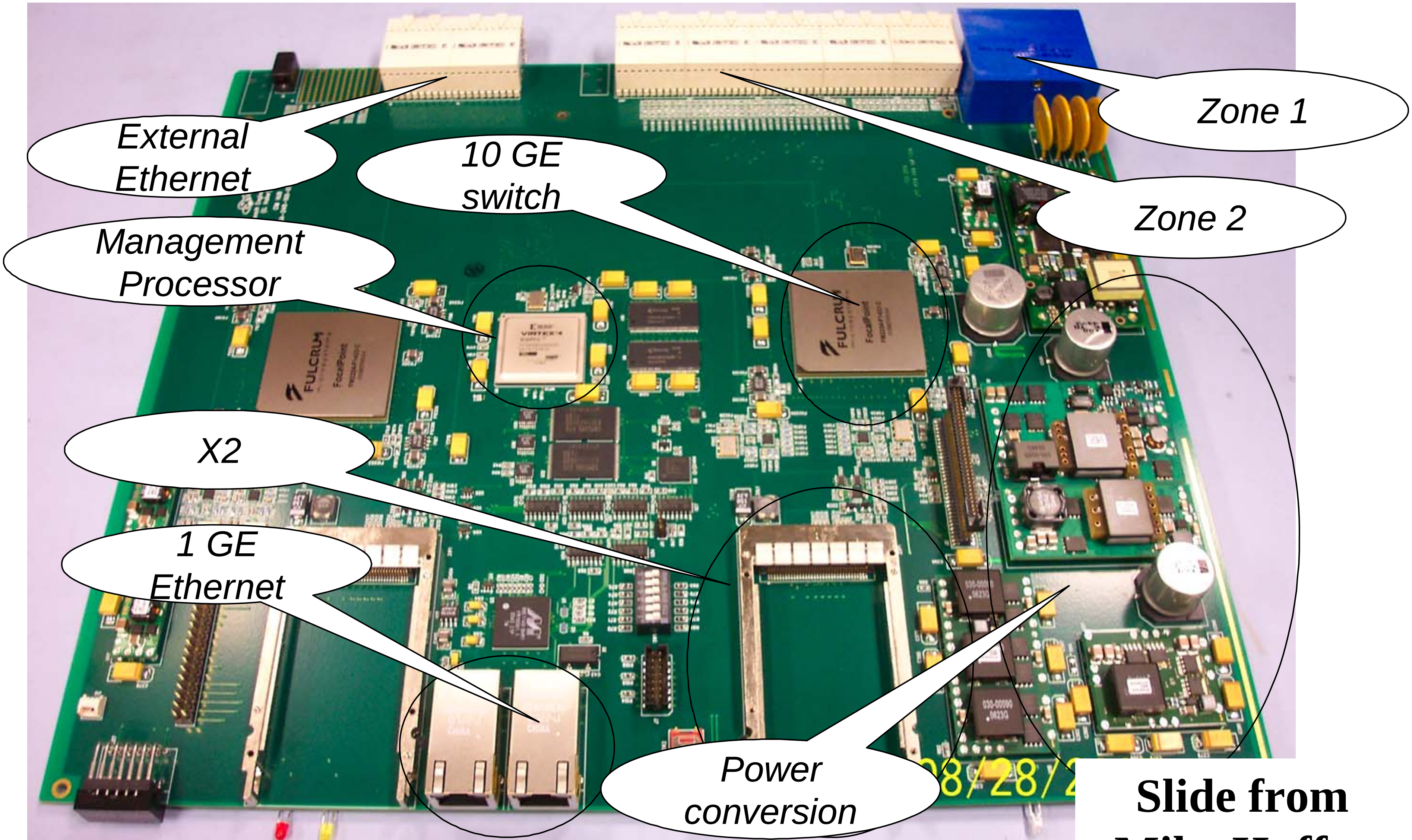
Media & Media Carrier

- Device is based on *Samsung* K9XXG08XX family
  - Nominally 4 GByte device
- Carrier is SO-DIMM
  - Contains 8 devices



Slide from Mike Huffer

Evaluation board



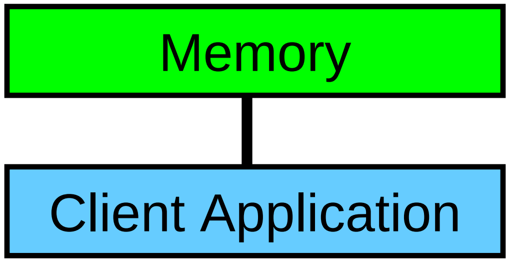
Slide from Mike Huffer

Commercial Product

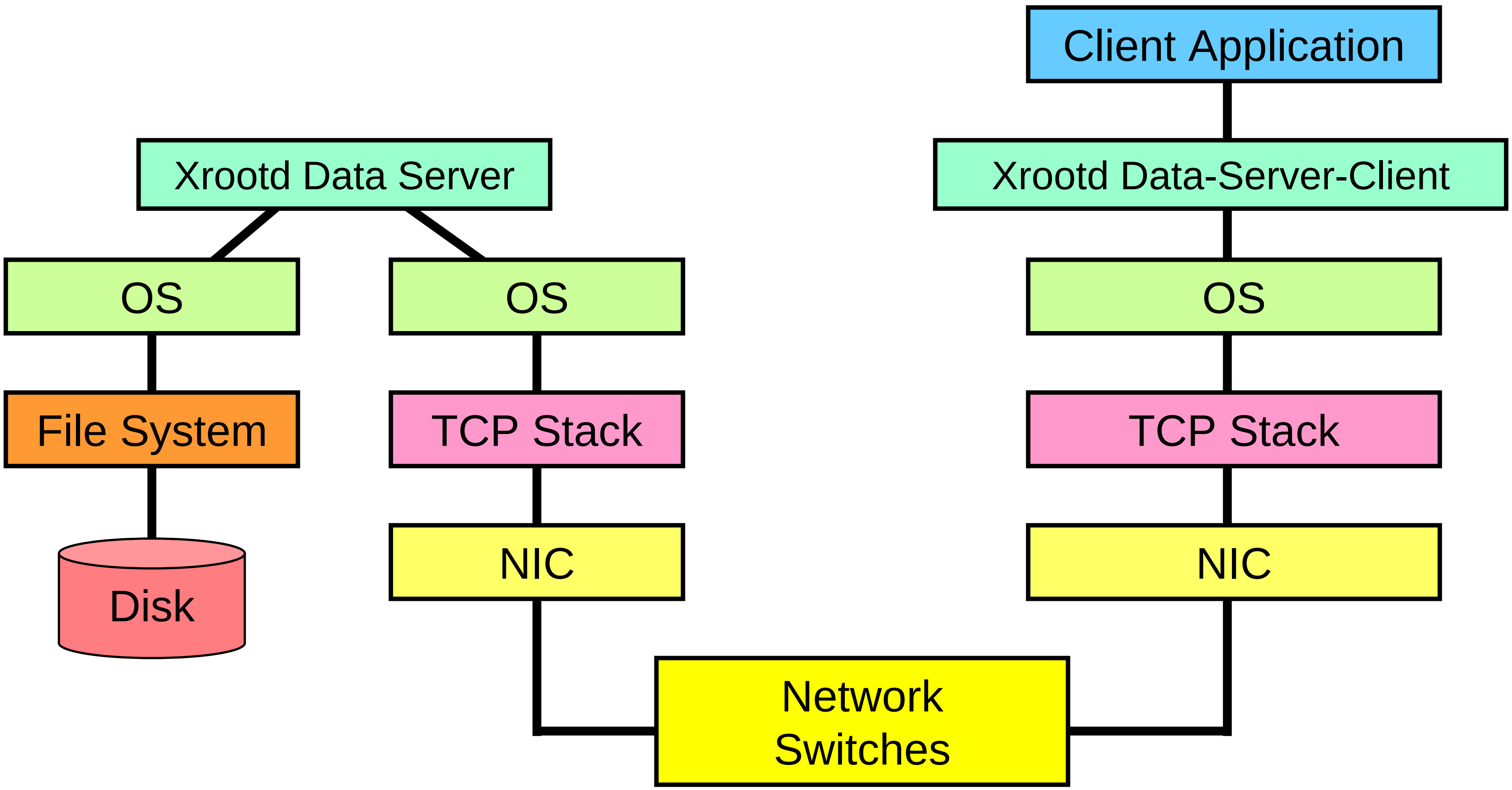
- Violin Technologies
  - 100s of GB of DRAM per box (available now)
  - TB of Flash per box (available real soon now)
  - PCIe hardware interface
  - Simple block-level device interface
  - DRAM prototype tested at SLAC

Some Performance Measurements

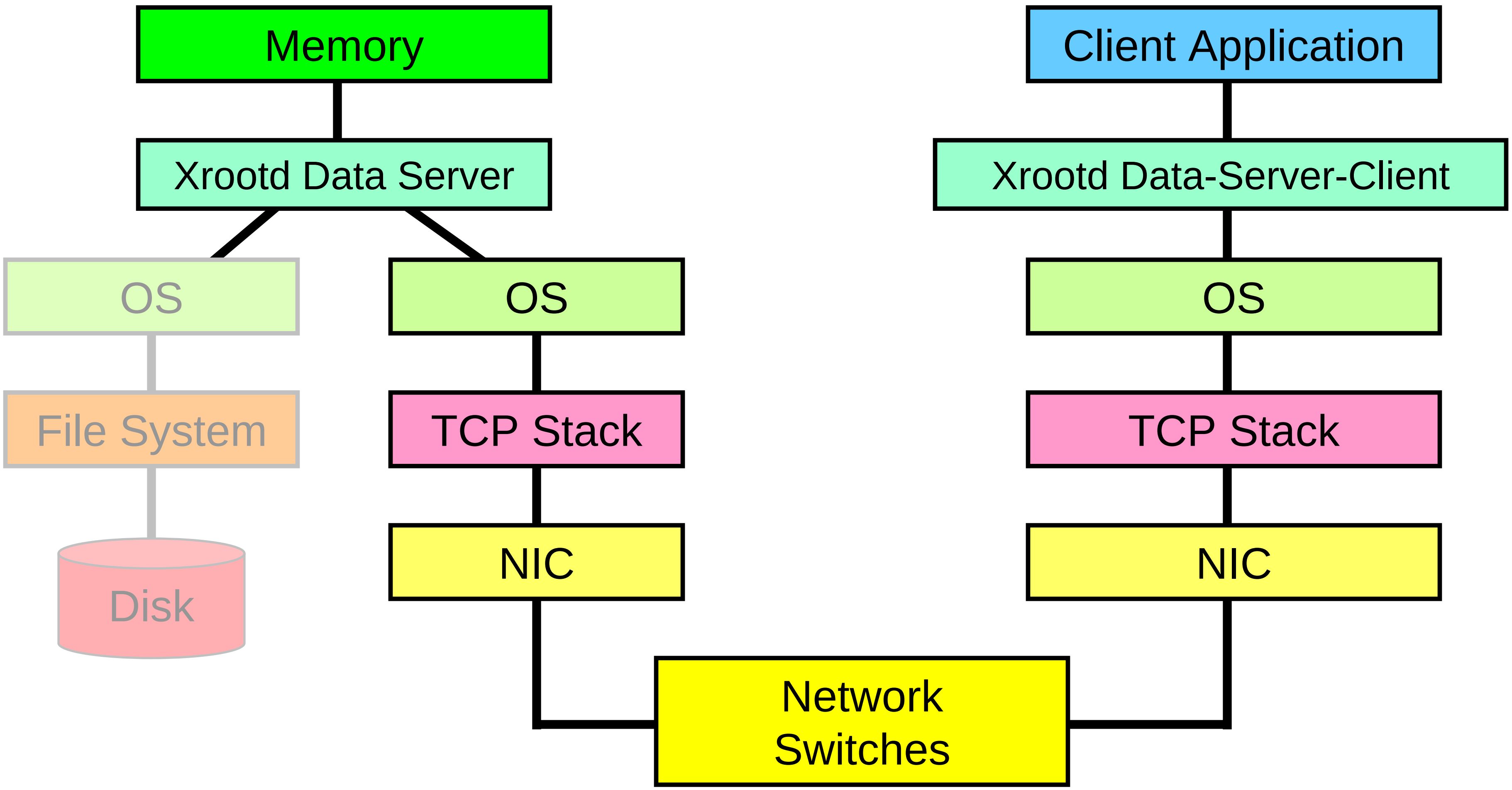
Latency (1)
Ideal



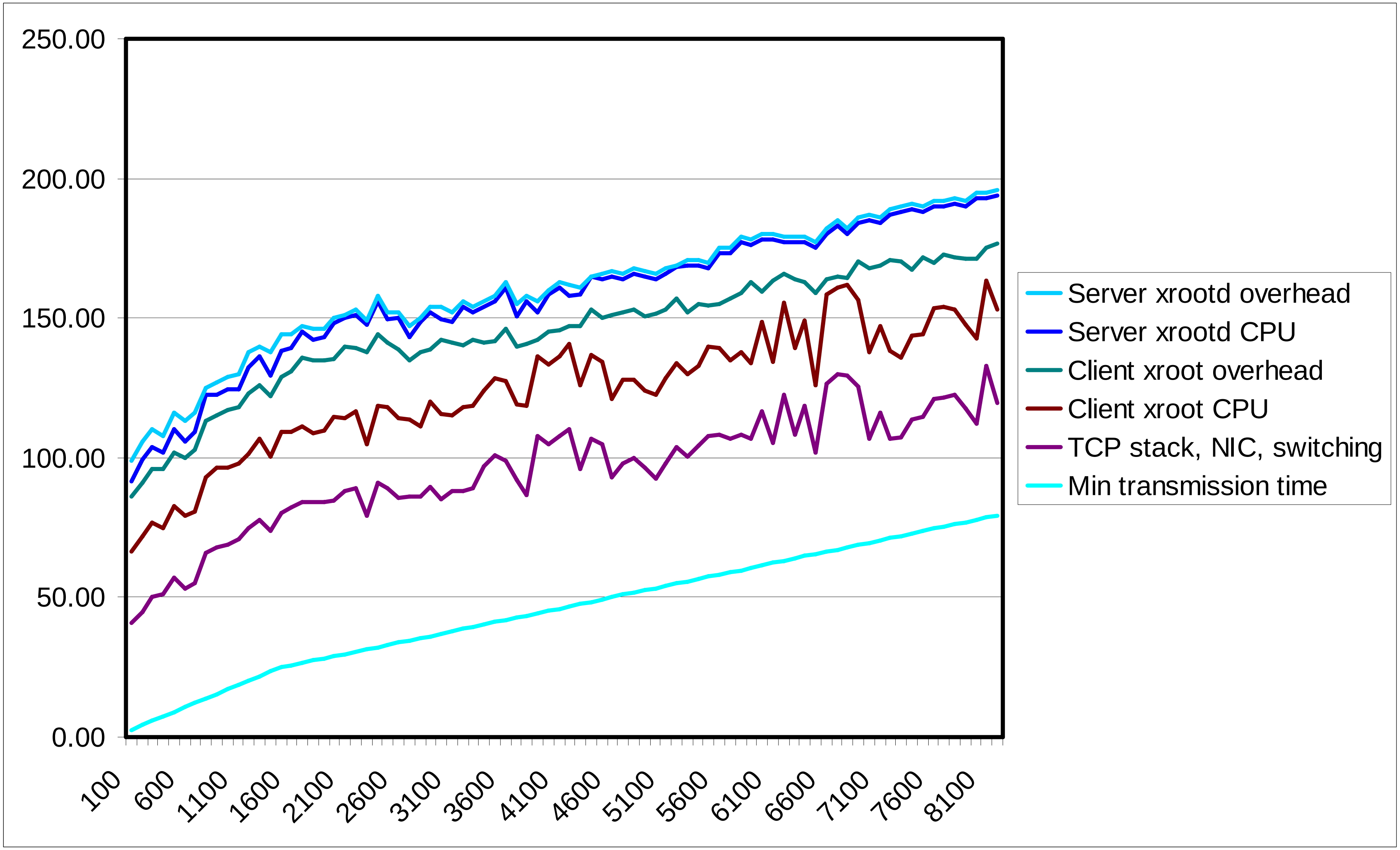
Latency (2)
Current reality



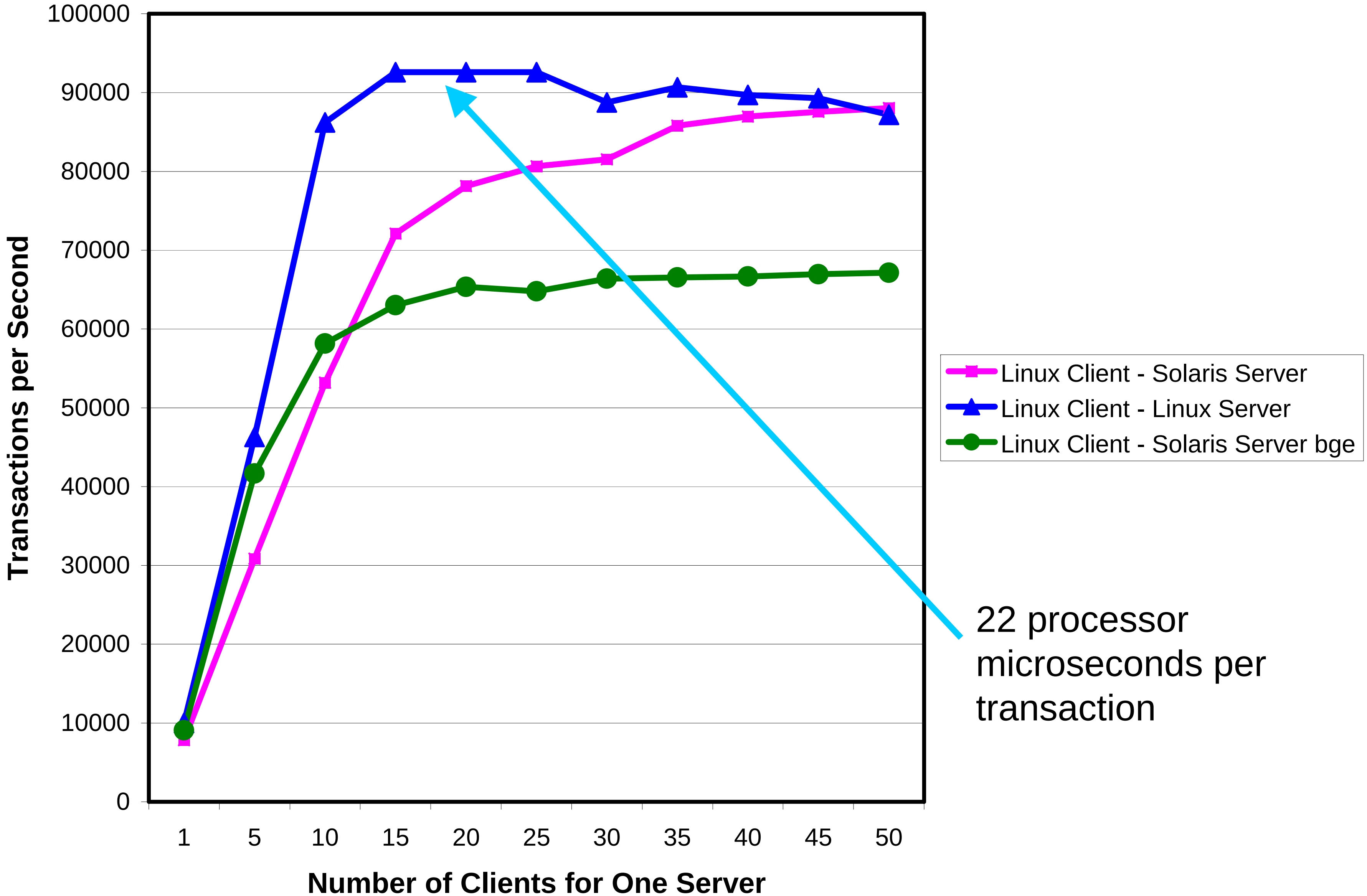
Latency (3)
Immediately Practical Goal



DRAM-Based Prototype
Latency (microseconds) versus data retrieved (bytes)



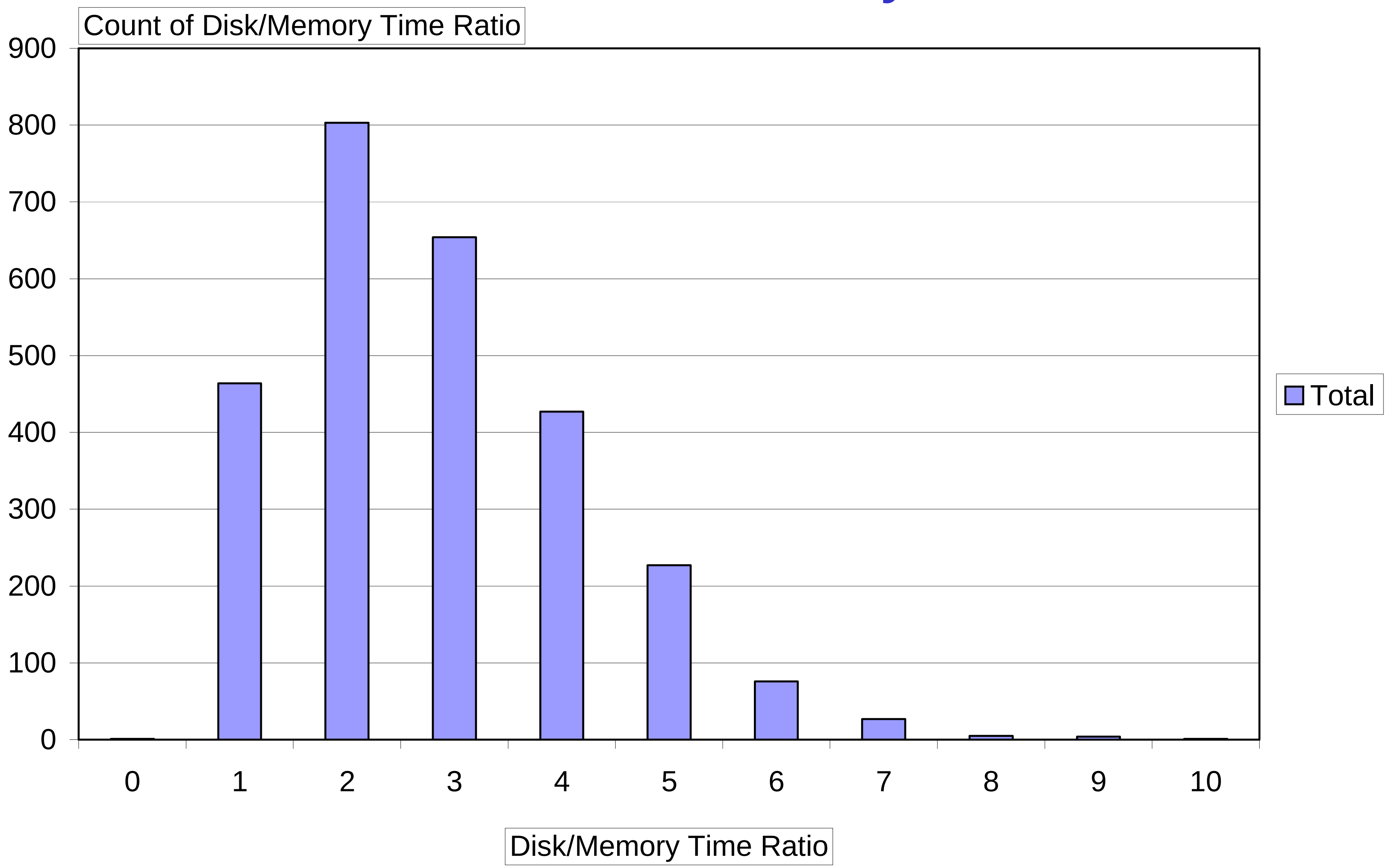
DRAM-Based Prototype Throughput Measurements



Throughput Tests

- ATLAS AOD Analysis
  - 1 GB file size (a disk can hold 500 – 1000 of these)
  - 59 xrootd client machines (up to 118 cores) performing top analysis getting data from 1 server.
  - The individual analysis jobs perform sequential access.
  - Compare time to completion when server uses its disk, compared with time taken when server uses its memory.

DRAM-based Protoype ATLAS AOD Analysis



Comments and Outlook

- Significant, but not revolutionary, benefits for high-load sequential data analysis – as expected.
- Revolutionary benefits expected for pointer-based data analysis – but not yet tested.
- The need to access storage in serial mode has become part of the culture of data-intensive science – why design a pointer-based analysis when its performance is certain to be abysmal?
- TAG database driven analysis?