Guided Wave Optics Lab
University of Colorado at Boulder

Optics Interconnections and Power Consumption in Cloud Computing

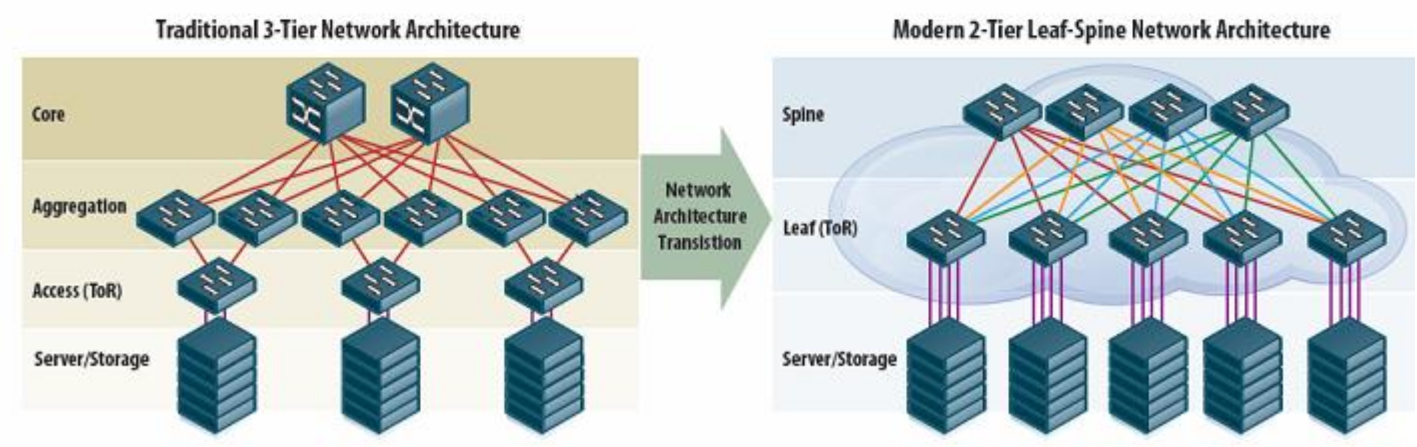
Moustafa Mohamed

Alan Mickelson

Guided Wave Optics Laboratory

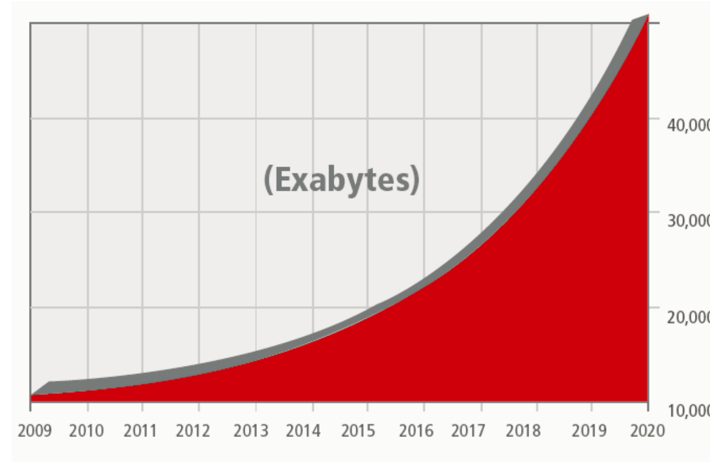
University of Colorado at Boulder

A Present Day Data Center



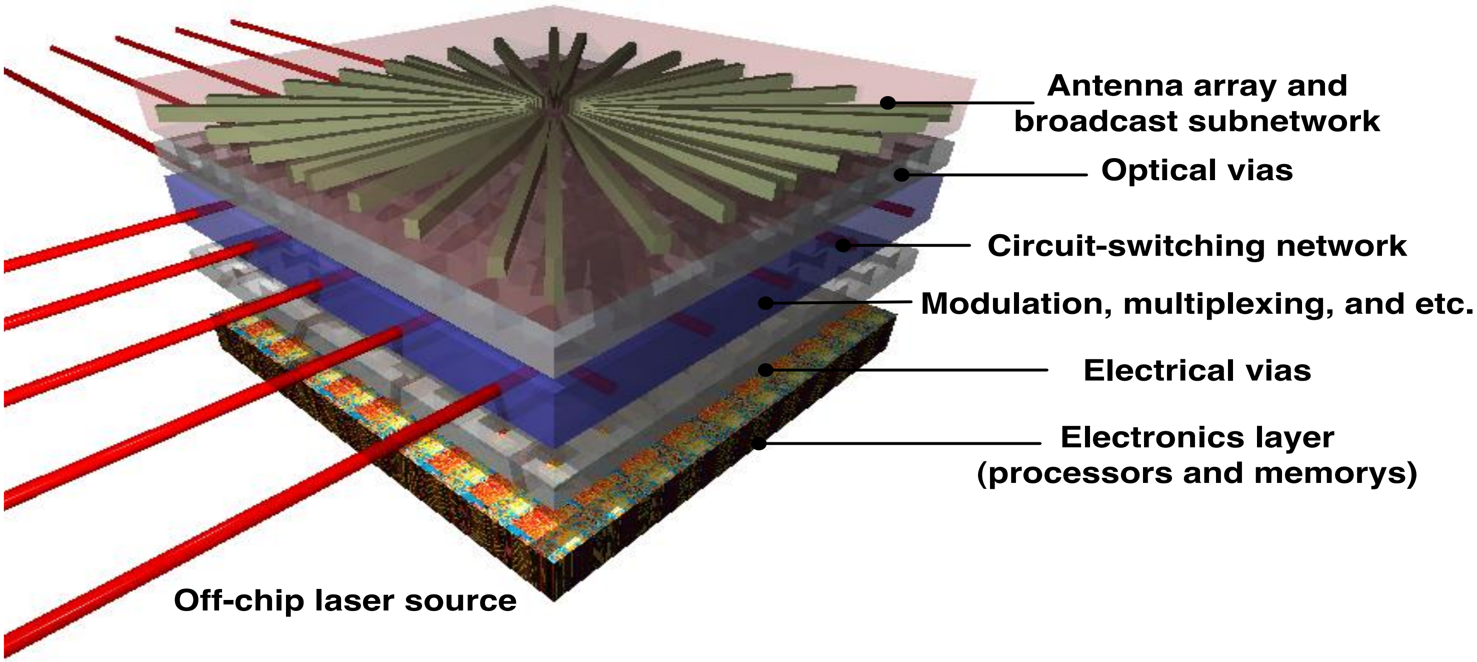
- The optical part is the network transceivers
- The interconnect is between GbE cards
- Inter rack is copper, Ethernet switches are electrical, transceivers are 10 Gbps (to be 100)

Data Center Energy Consumption (true but somewhat misleading)



- 91 TW-h of electricity in 2013 (140 TW-h in 2020) - World total is 25 PW-h (roughly 0.4%)
- Storage requirements grow up to 40%/year
- Information doubles every 18 – 24 months
- Best present metric of data center energy scaling is the storage requirement
- Present storage is roughly 100 P Bytes

Research Question

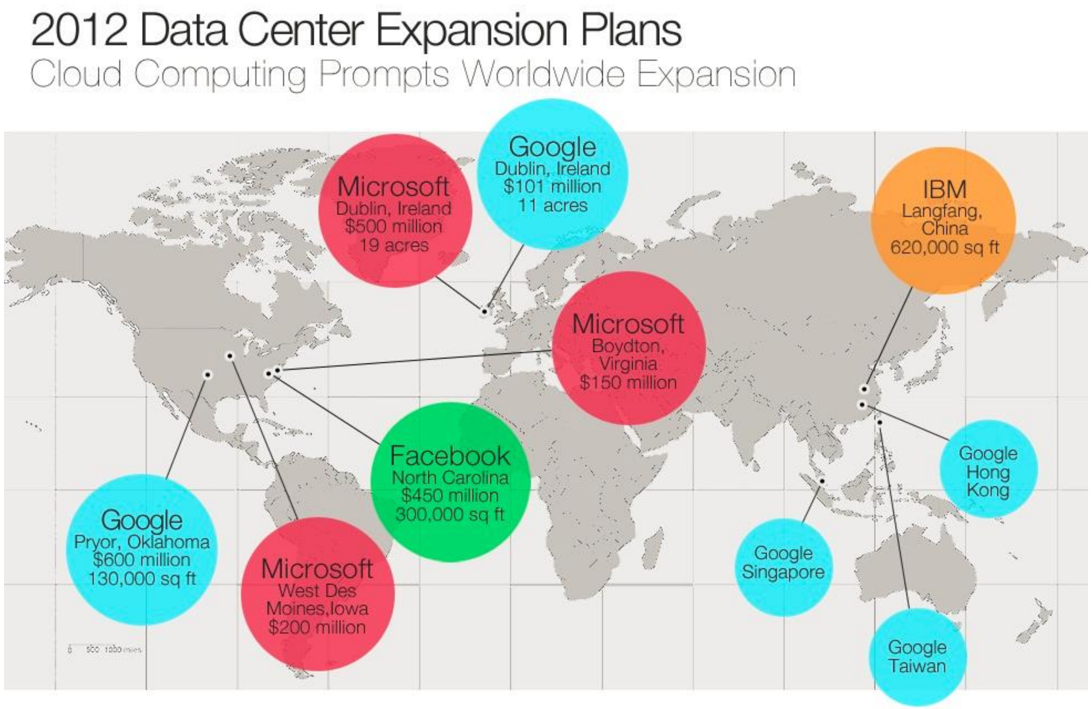


- Where should optics be in the Data Center?
- Rack-top, inter-rack, intraboard or chip level?

Outline

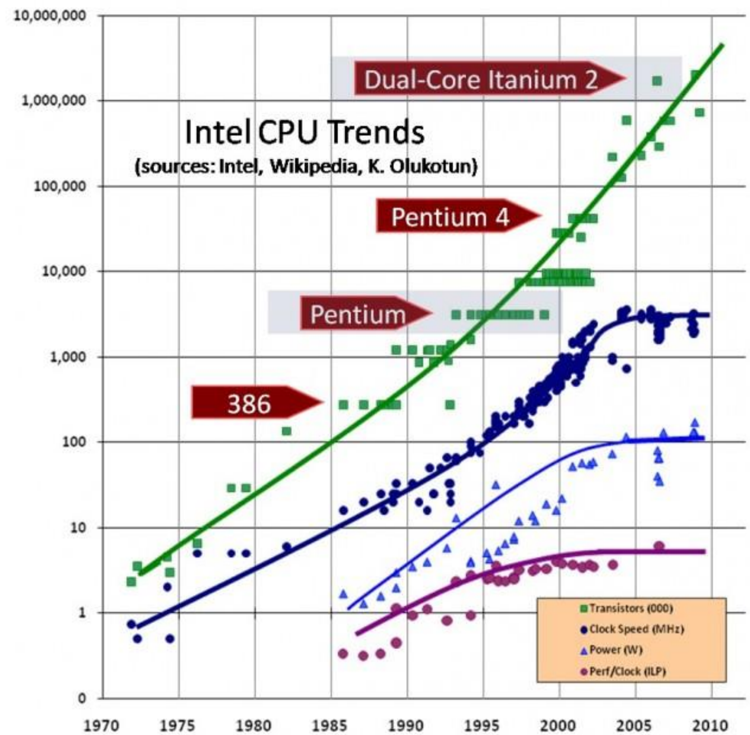
- The Problem
- Optical Interconnections
- Trends in Optical Interconnection
- Trends in Efficiency in Supercomputing
- Where Do Improvements Come From
- How to Evaluate Optical Interconnections
- A Simulation Framework
- Some Simulation Results

The Problem



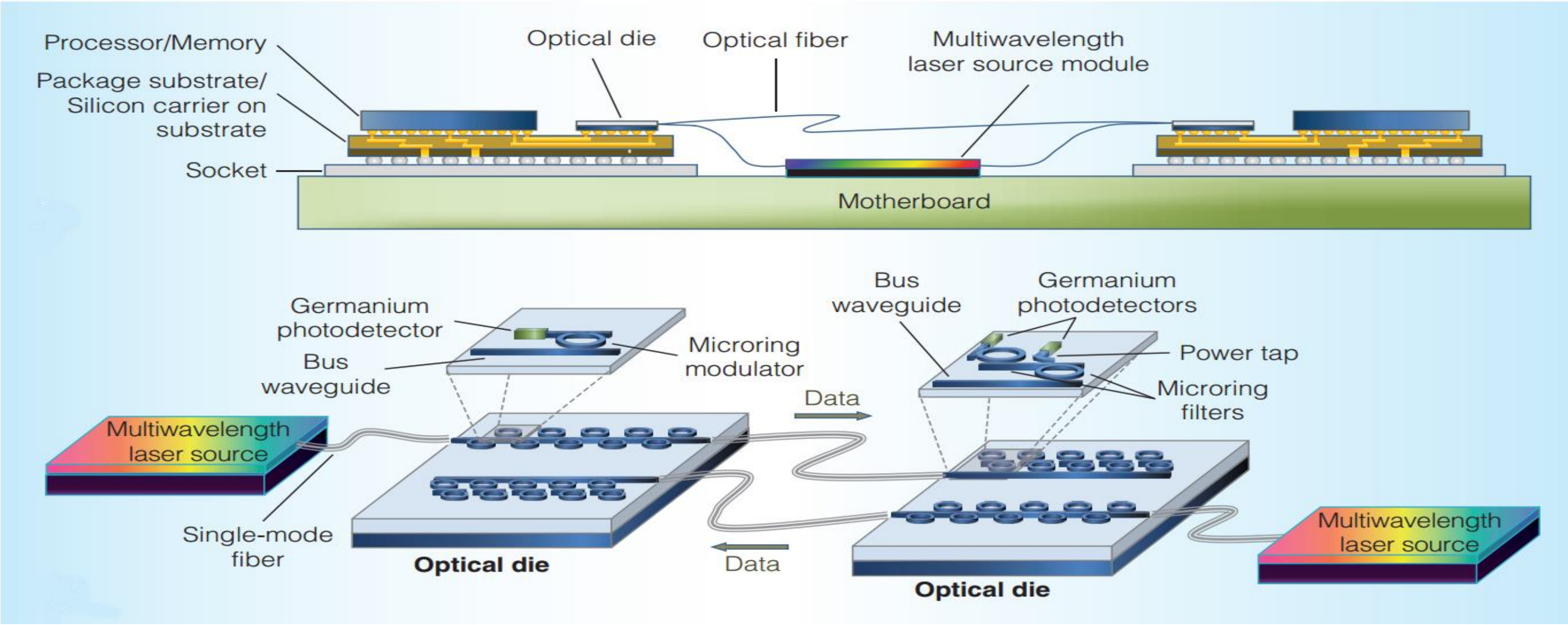
- Data center capacity will increase to the level of an Exa-Byte in the near future
- Drivers for further growth are strong

The Problem



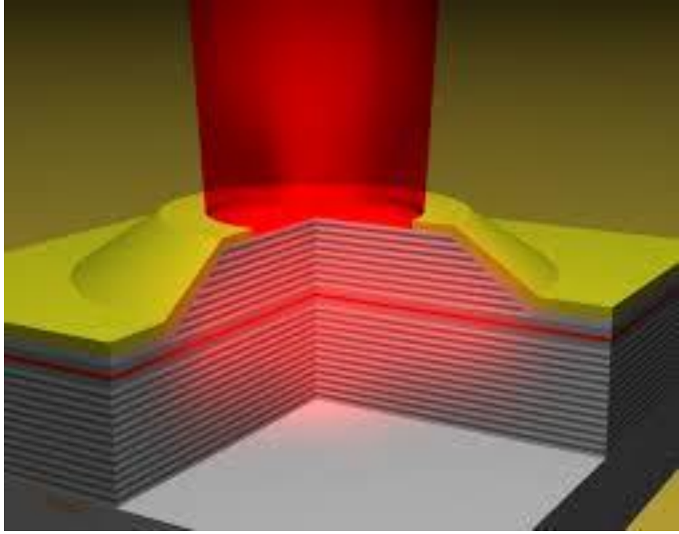
- **Moore’s law can no longer be counted on to support exponential improvements**

Optical Interconnections



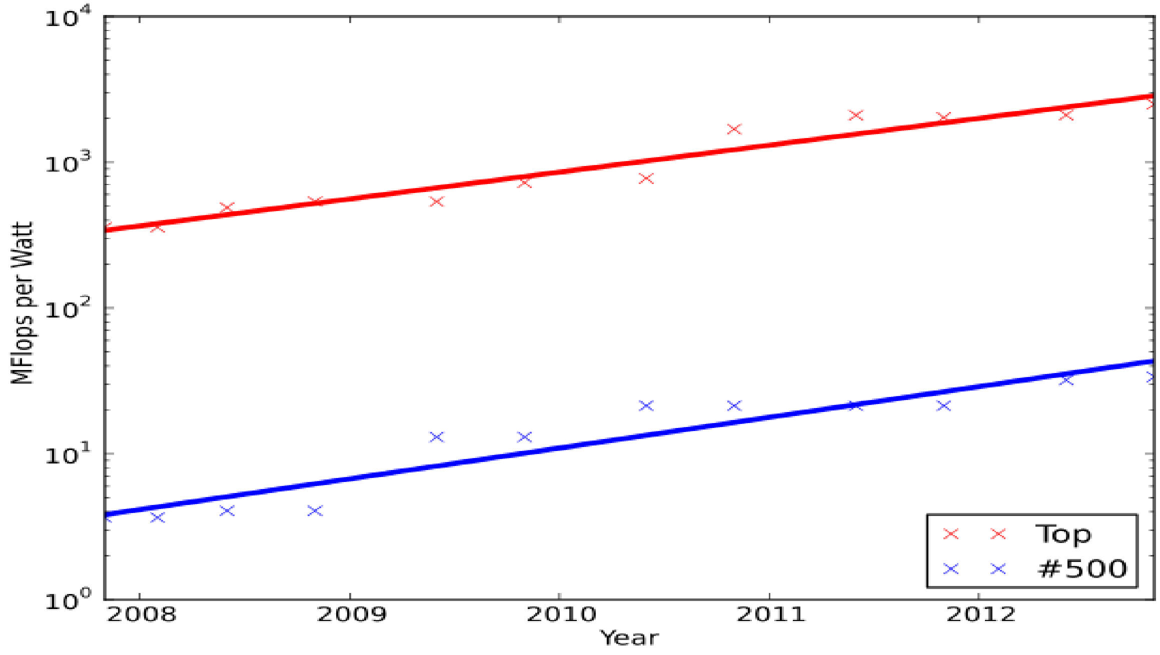
- **Silicon photonics offers optical advantages at ever smaller interconnect distances (figure from Columbia – taken from Web)**

Trends for Optics In Computing 2000 - 2015



- 1 Gbps VCSEL modules became available in 2000 and rapidly became data com standard
- 10 Gbps VCSELs modules became available in circa 2005 and are present state of the art
- With leaf spine architecture, 10 Gbps VCSELs connect 10 GbE boards
- Some 4 x 10 Gbps connections have become available recently to connect 10 GbE boards

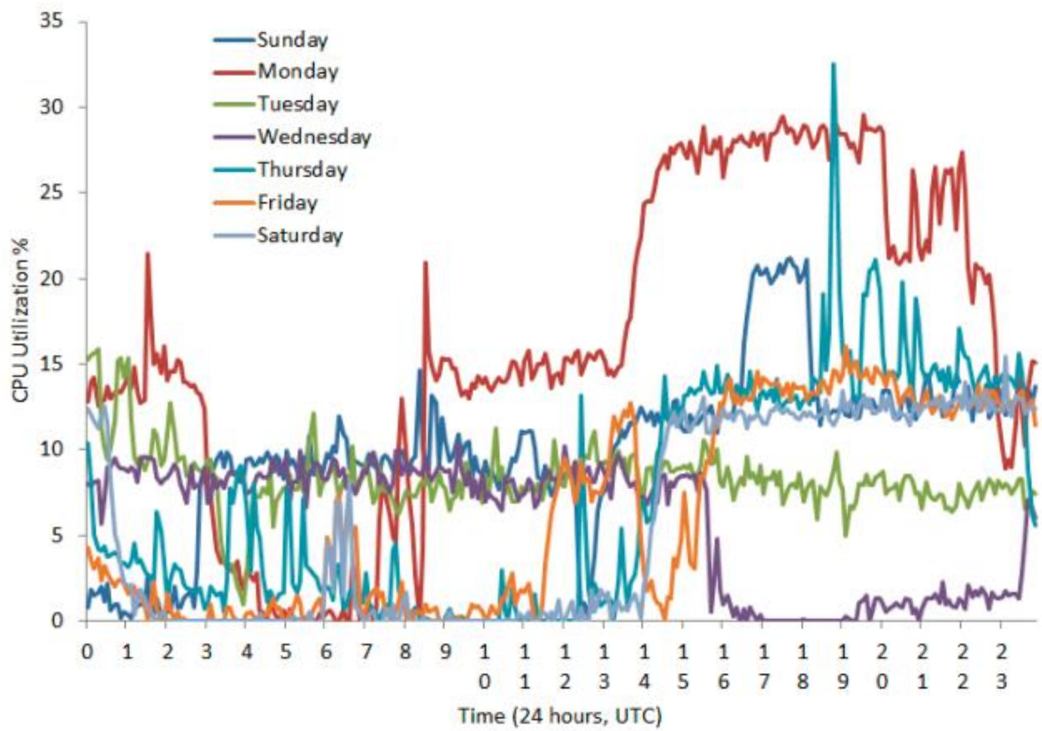
Efficiency in Supercomputing



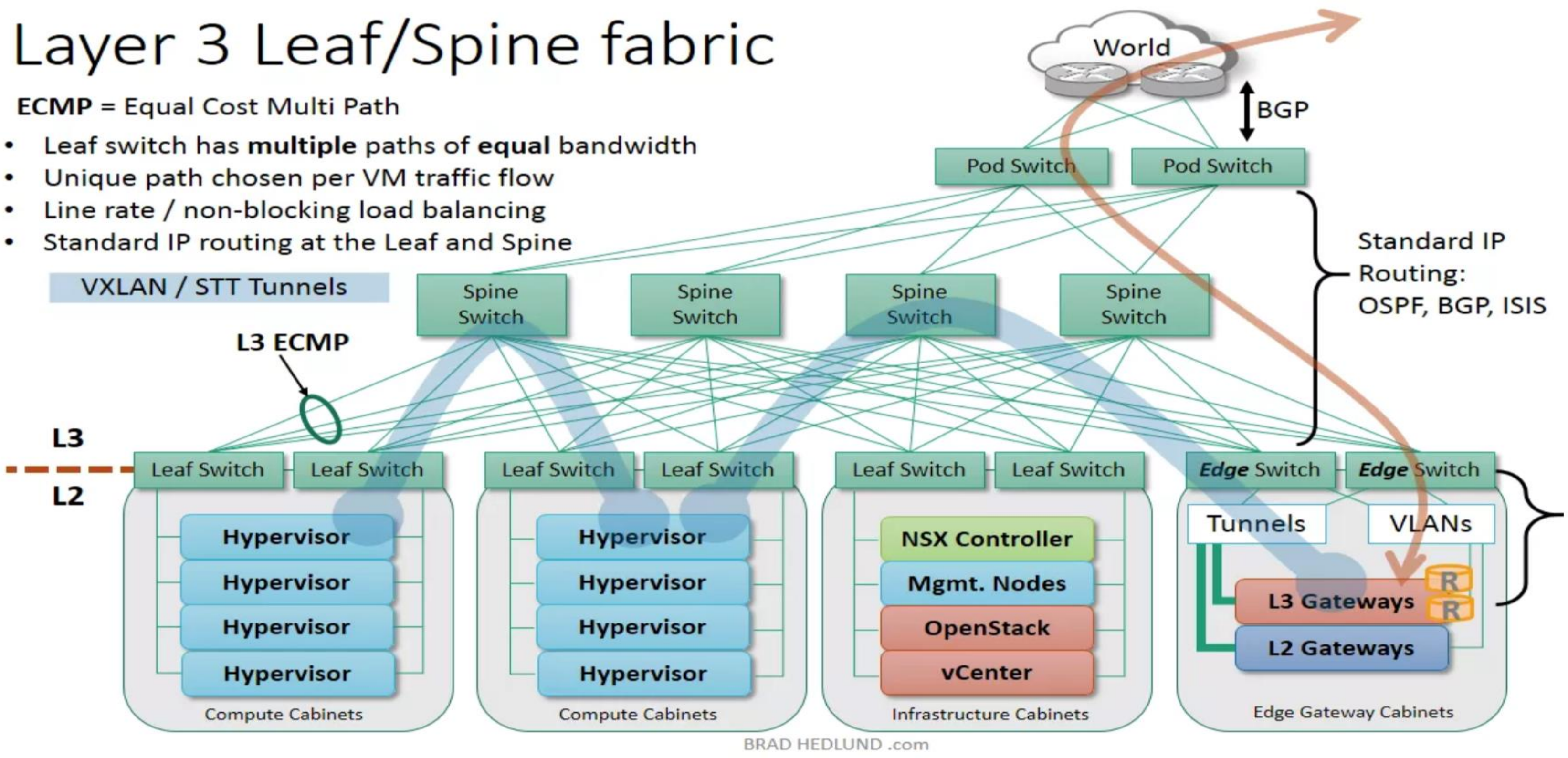
- **Supercomputer power efficiency has increased 1000 since 2005 with no change in the optical interconnections**

Where Do Improvements Arise

- Blade servers have improved significantly
- Idling servers draw power – no energy proportionality in data centers



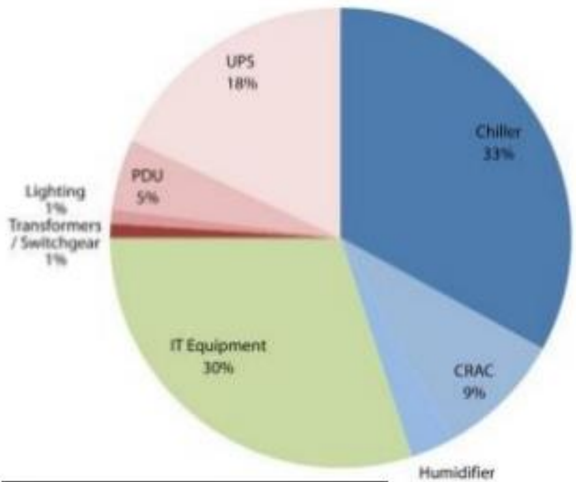
Where Do Improvements Arise



- Virtualizing interconnections in leaf spine can level loads and eliminate queues

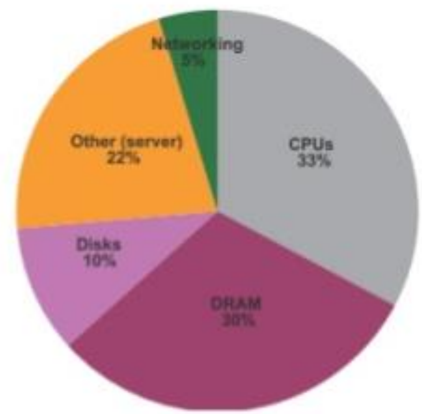
Optical Interconnections

Data Center Energy breakdown



Source: ASHRAE

Server Peak power usage profile



Source: Google 2007

- The interconnection including optics, GbE board and all drivers uses 1.5% of DC energy
- Speed increase will cause energy increase
- Savings must come from load leveling/other

Simulating Data Center Performance

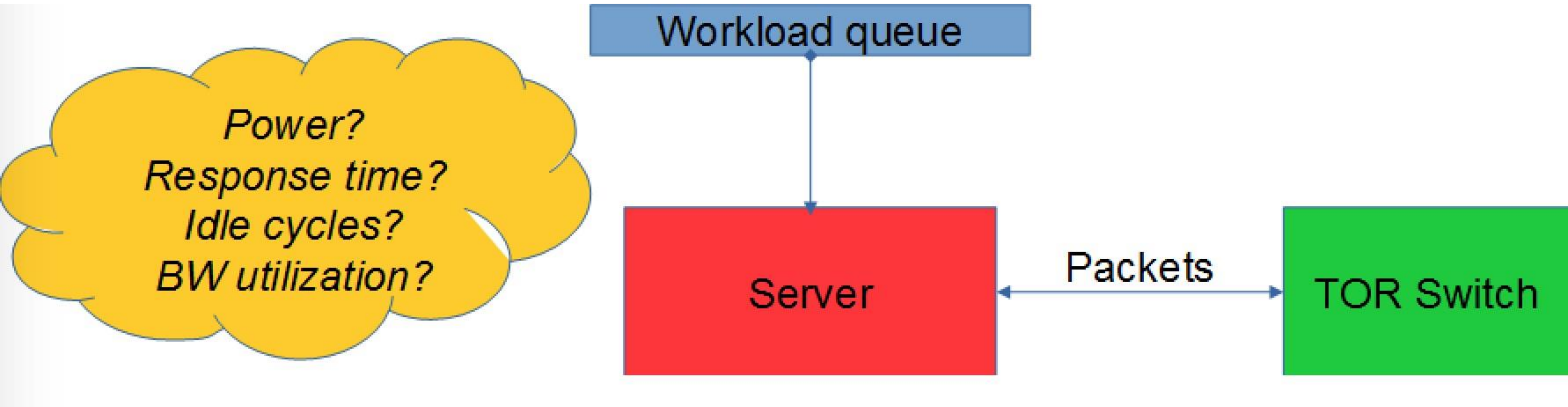
$$\mathbf{p_n(n) = \frac{\lambda^n}{n!} e^{-\lambda}}$$

$$\mathbf{\bar{n} = \lambda} \qquad \mathbf{M/M/1 \leftrightarrow M/G/K}$$

$$\mathbf{\sigma_n = \lambda}$$

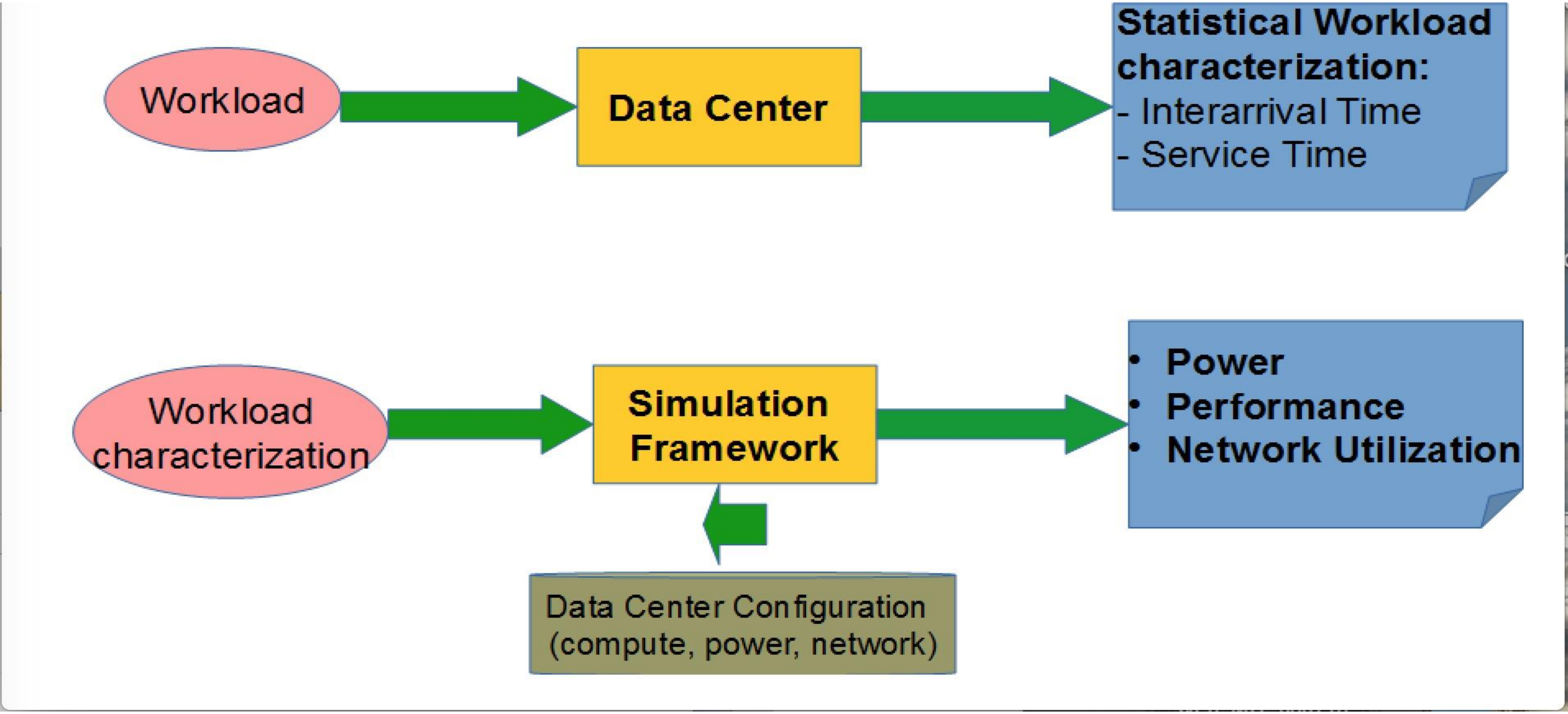
- **Center traffic is heterogeneous and bursty**
- **Poor usage results from queues – traffic does not share equally among the cores**
- **A statistical approach (even that ignores traffic correlation) can be advantageous**

Simulating Data Center Performance



- **A University of Michigan open source tool – Big House – is available – we are building**

Simulating Data Center Performance



- **The tool can predict power usage as well as the statistics of performance parameters**

Some Measurement Characteristics

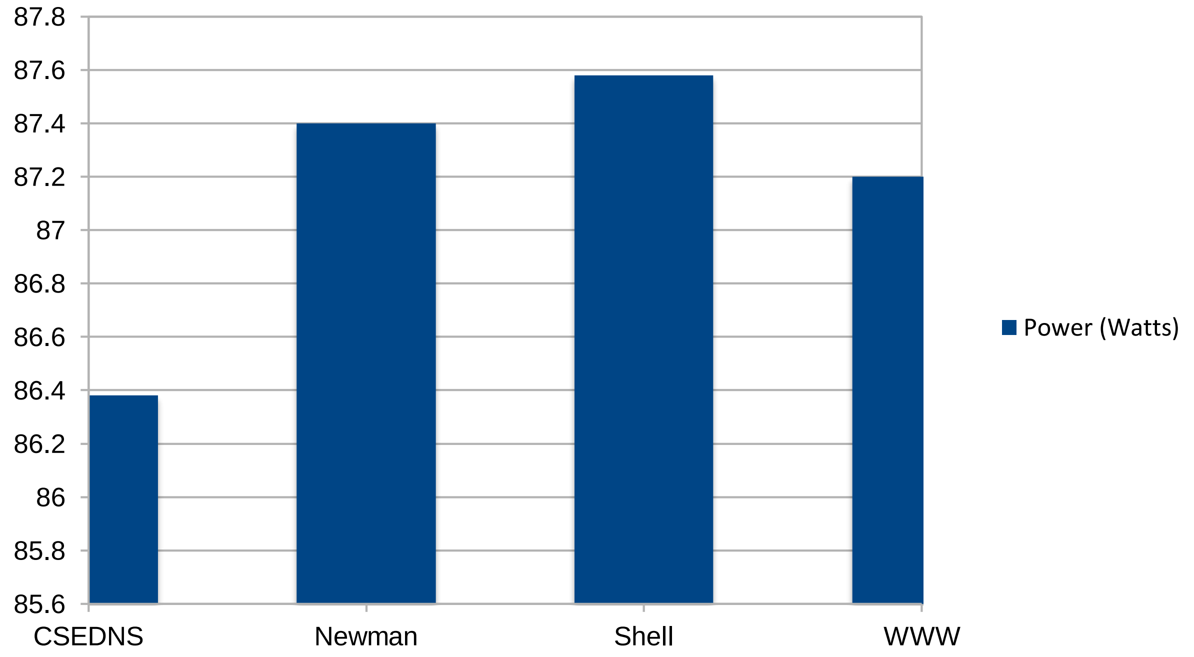
Workload	Interarrival			Service		
	Avg	σ	C_v	Avg	σ	C_v
DNS	1.1 s	1.2 s	1.1	194 ms	198 ms	1.0
Mail	206 ms	397 ms	1.9	92 ms	335 ms	3.6
Shell	186 ms	796 ms	4.2	46 ms	725 ms	15
Web	186 ms	380 ms	2.0	75 ms	263 ms	3.3

- **For a Poisson interarrival with homogenous traffic, Cv would be one**
- **Typical data center traffic is high Cv (up to 100)**

Workload Definitions

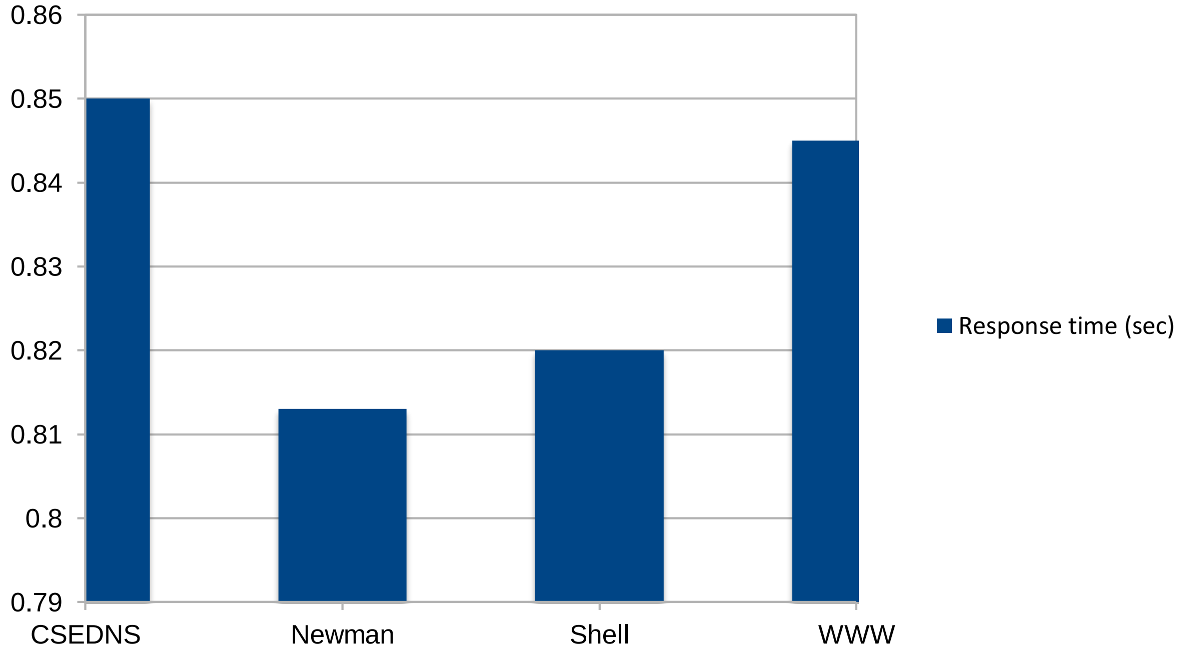
- DNS refers to Domain Name Server and Dynamic Host Configuration Protocol Server under live traffic
- Mail refers to Post Office Protocol 3 and Simple Mail Transfer Protocol server under live traffic
- Shell refers to login server under live traffic
- Web refers to Hypertext Transfer Protocol server under live traffic

Simulation Results: Power



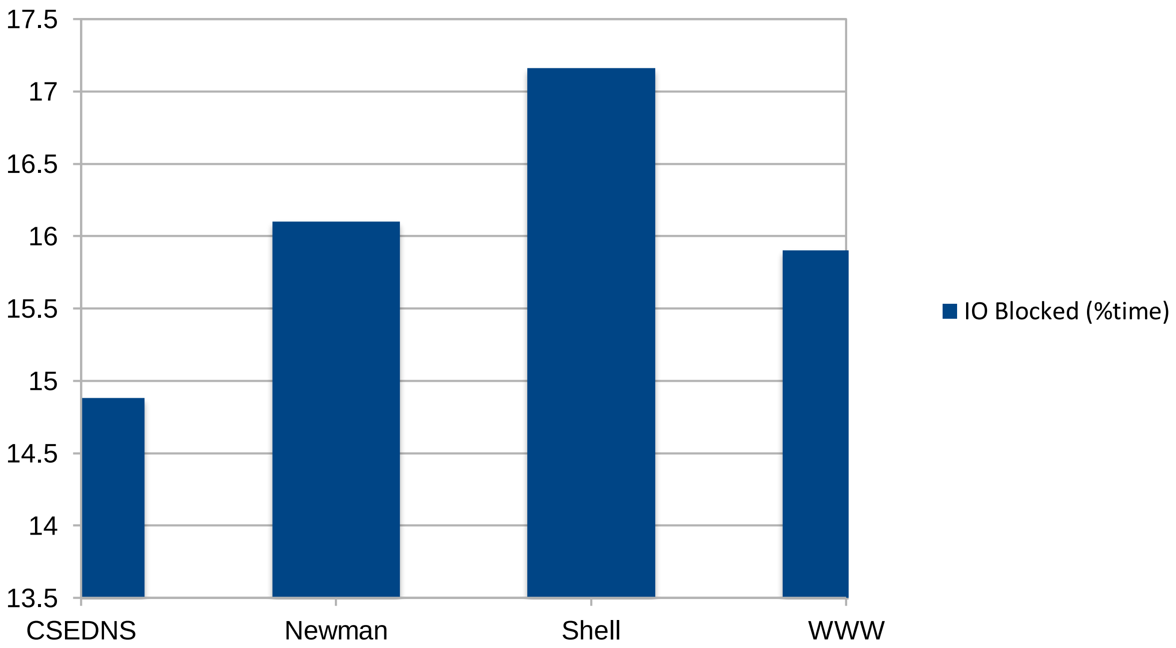
- **All simulations are for 10,000 servers operating at 10 Gbps**
- **Servers are using DVFS to conserve power**

Simulation Results: Response Time



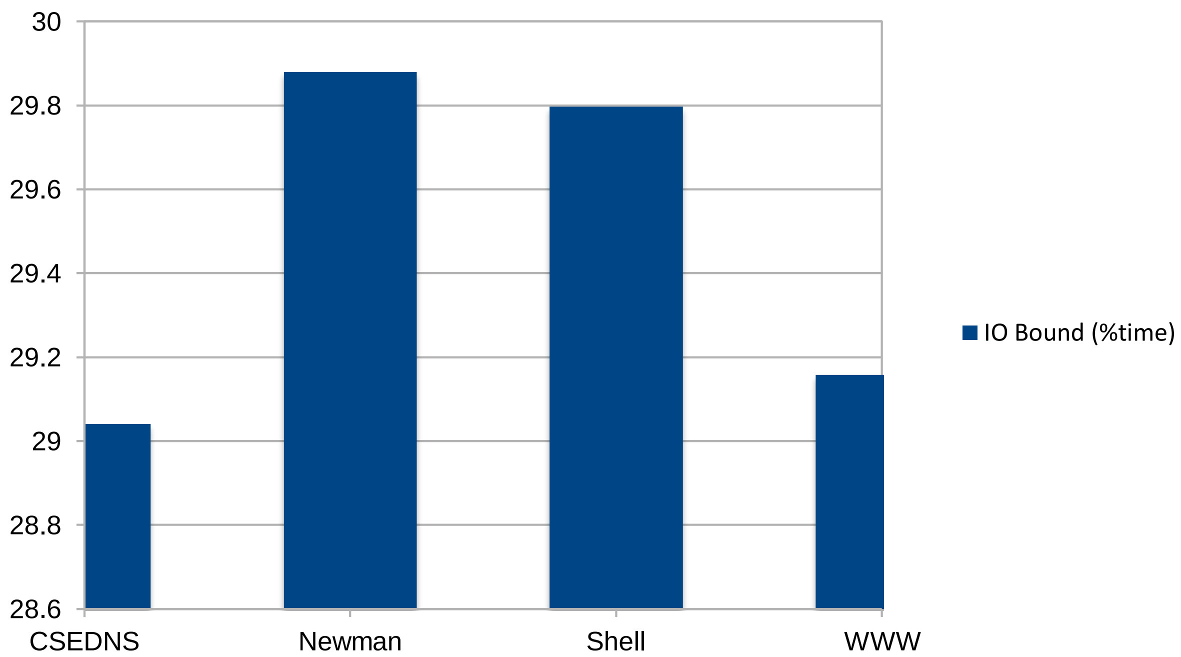
- **Response time is between arrival and job completion**

Simulation Results: IO Blocking



- **Percentage of time the CPU was limiting operation**

Simulation Results: IO Bound Time



- **Percentage of time that the interconnect was limiting delivery to CPU's**

Summary

- Replacing copper with fiber increases data center throughput but not proportionally
- Power efficiency is tied to uniformity of the computational load across the data center
- A simulation tool, Big House, based on queuing offers insight into load leveling
- Initial simulations yield well known results on efficiency versus power and bandwidth

Questions