Data Center Fundamentals:
The Datacenter as a Computer

George Porter
CSE 291
April 2, 2015

*Includes material taken from Barroso et al., 2013, UCSD 222a,
and Cedric Lam and Hong Liu (Google)
Traditional DC Topology

Core

Aggregation

Access

Layer-3 router

Layer-2/3 switch

Layer-2 switch

Servers
DC Network Requirements

• Scalability
  – Incremental build out?

• Reliability
  – Loop free forwarding

• VM migration

• Reasonable management burden
  – Humans in the loop?
Traditional Topologies

• Over subscription of links higher up in the topology

• Tradeoff between cost and provisioning

• Single point of failure
Capacity Bottlenecks
Building block: \( N \times N \) Non-Blocking Switch

- Requires \( N^2 \) cross-bar switches
- High-radix (port count) 10G-and-above line-rate \( N \times N \) switching silicon (a few 10s of ports) are readily available from silicon vendors
  - Building blocks for large scale multi-stage switching fabrics
Cluster Topology: Torus

- Mesh connection with every row and column wrapped around as a ring
- 2-D torus is like a donut
  - Reduces radius by $\frac{1}{2}$ to $\sqrt{N}$
Cluster Topology: Hyper Cube

- Multi-dimensional cubes
- $n$ dimensions, $N = 2^n$ nodes
- Nodes identified by $n$-bit numbers
- Each node connected to other nodes that differ in a single bit
- Degree: $\log_2 N$
- Diameter: $\log_2 N = \text{dimension}$

Recursive construction of a 4-D hyper cube

Switch Chips

Small-Radix Router

- Simpler internal switching logic
- Lower digital switching power
- Lower serial latency
- Higher header latency
- Higher cost

High-Radix Router

- Complex internal switching logic
- Higher digital switching power
- Higher serial latency
- Lower header latency
- Lower cost

- High radix switches are important to support network fabric with much larger bisectional bandwidth and contain cost
  - E.g., Clos with 16x40Gb/s versus 64x10Gb/s building blocks in 3-stage, 4Tb/s versus 64Tb/s max bisection bandwidth
Switch I/O bandwidth is scaling one decade per 5 years.

- Trade-off between port bandwidth and port count
- Switch port counts limited by pin counts and I/O power

- High speed SerDes are widely used to increase bandwidth
- Transmitter pre-emphasis and receiver equalization compensate the physical channel impairment (ISI, xtalk, etc.)
System Packaging: Chassis Capacity

- The chassis capacity is limited by front-panel I/O density, backplane capacity and power

- Front-panel I/O density
  - Line card configuration
    - 48 ports of SFP+ @ 10Gb/s = 480 Gb/s
    - 40 ports of QSFP @ 40Gb/s = 1.60 Tb/s
    - 16 ports of CXP @ 120Gb/s = 1.92 Tb/s
  - Required backplane capacity
    - 8 line cards: 3.84 Tb/s to 15.36 Tb/s
    - 16 line cards: 7.68 Tb/s to 30.72 Tb/s

- Power also needs to scale to meet ever increasing chassis capacity
System Packaging: High Speed Backplane

- **IEEE P802.3ap – Ethernet Backplane**
  - Support up to 1m trace on improved FR-4 PCB + 2 connectors
    - 10GBASE-KR: 10Gb/s serial
      - 64B/66B encoding
      - TX pre-emphasis, adaptive RX equalization + DFE, FFE, FEC (3dB)
    - 100GBASE-KR4: 25 Gb/s over 4 lanes
      - CAUI 25Gb/s/lane
      - 64B/66B encoding
      - TX pre-emphasis
      - Adaptive RX equalization + DFE, FFE, FEC (5dB)

- **Key considerations**
  - Improved design practices for higher speeds
  - Better board material: PCB trace is dielectric loss limited
  - Better SI practices: connector and pin-out, vias, back drill, etc.
Optical Backplane?

- High speed transmitters TX
- High speed receivers Rx
- Embedded low-loss waveguide on boards
- In-plane optical turns
- Out-of-plane optical turns
- Optical connectors for light coupling
- Optical jumpers

- Optics plays different role in the last 1m.
- How can optics deliver a power efficient backplane solution at the cost of copper PCB?
System Packaging: Copper Cables

- Cu cable provides lower cost, short reach solution for intra-rack
- Leverage backplane SerDes technology and standards (10GBase-CR, 40GBase-CR4, 100GBase-CR4, 100GBase-CR10, etc.)
- The limiting factors are:
  - Copper loss (dominant loss)
  - Dielectric loss (2\textsuperscript{nd} order loss)
  - Pair shield, different skew.
- Cable bulk and weight challenges for high density system
System Packaging: Copper Cables

- Electrical dispersion compensation continues to drive copper cable to higher data rate and longer reach.
- But it is increasingly difficult and costly ($, power and density).

10Gbps Electrical Link in 90nm CMOS over 1m, 5m of 26 Gauge Cable

No Equalization, 1m

Equalization, 5m

Cable + PCB loss: -10dB at 5GHz

Cable + PCB loss: -20dB at 5GHz

Transmitter has one pre, one post tap
Single-Mode vs Multi-Mode Fiber

- **Single-Mode Fiber**
  - More bandwidth-distance
  - Better support for WDM
  - Tighter connector tolerance

- **Current VCSEL transmitters do not operate with SMF**
  - Edge emitting lasers are more expensive and consumes more power

- **Have the potential as single unifying media for both intra-datacenter and inter-datacenter transmissions**

- **Multi-Mode Fiber**
  - Limited bandwidth-distance
  - Large connector tolerance
    - Not a major advantage for datacenter

- **Cheaper and lower power VCSEL based transmitters**

- **Parallel optics implementation requires ribbon fiber cables and connectors**
  - Expensive cable solution
Background: Fibers

**Single Mode Fiber**

**Multi Mode Fiber**

SMF: core diameter: 9\(\mu\)m, MMF: core diameters 50\(\mu\)m, 62.5\(\mu\)m

- **MMF (Multi Mode Fiber) has limited bandwidth and more expensive**
  - Modal dispersion limited: Multiple modes propagate at different velocities, causes pulse spreading
  - Graded parabolic index profile reduces modal dispersion
    - requires careful control of doping concentration layer by layer from the center of the core

- **SMF (Single Mode Fiber) has Tb/s bandwidth and less expensive**
## Metric I: Cabling Simplicity

<table>
<thead>
<tr>
<th></th>
<th>SMF Cable</th>
<th>MMF Cable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost</td>
<td>Volume</td>
</tr>
<tr>
<td>10G (4x10Gb/s)</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>40G (4x25Gb/s)</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>100G (4x25Gb/s)</td>
<td>1x</td>
<td>1x</td>
</tr>
<tr>
<td>400G (16x25Gb/s)</td>
<td>1x</td>
<td>1x</td>
</tr>
</tbody>
</table>

![Cable Diagrams](image_url)
Challenges for VCSEL as high speed transmitter

- Low output power due to short gain region -> limited reach of 100m for 40G/100G array, tight power budget for patch panel and MTP
- Trade-off between power and speed: 25Gb/s is pushing the limit of oxide aperture size
The MBD module, using an LGA socket design, can be placed on a small 1-inch PCB grid.
Parallel Transceivers, Ribbon Cables, and AOC

- QSFP and CXP modules are used in Infiniband interconnect and 40GE/100GE
- Use multi-mode ribbon cables
  - MTP/MPO connector
    - 12-fiber version (QSFP)
    - 24-fiber version (CXP)
  - Cables and connectors are expensive
- AOC (active optic cable)
  - Optical interface is not exposed
  - Alleviates optical interface interoperability issue
  - Improves yield and power consumption
  - AOC cables are difficult to operate and manage
  - Hard to thread transceiver modules across fiber guides and conduits
# 40GbE and 100GbE Standards - IEEE

## IEEE Standards Reaches for 40GbE (optical)

<table>
<thead>
<tr>
<th>Name</th>
<th>Comments</th>
<th>Reach</th>
<th>Wavelength/Source</th>
<th>Form-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>40GBase-SR4</td>
<td>4 lasers each operating at 10 Gb/s. Electrical I/O is non-retimed 10 Gb/s.</td>
<td>100m (OM3) 125m (OM4)</td>
<td>850nm VCSEL</td>
<td>QSFP</td>
</tr>
<tr>
<td>40GBase-LR4</td>
<td>4 lasers each operating at 10 Gb/s. Electrical I/O is non-retimed 10 Gb/s.</td>
<td>10 km SMF</td>
<td>CWDM 1270,1290,1310,1330 nm</td>
<td>QSFP</td>
</tr>
<tr>
<td>40GBase-ER4</td>
<td>4 lasers each operating at 10 Gb/s. Electrical I/O is non-retimed 10 Gb/s.</td>
<td>40 km SMF</td>
<td>CWDM 1270,1290,1310,1330 nm</td>
<td>QSFP</td>
</tr>
</tbody>
</table>

![QSFP 40G SR4](image1.png)  ![QSFP 40G LR4](image2.png)
## 100GbE Standards - IEEE

### IEEE Standards Reaches for 100GbE (optical)

<table>
<thead>
<tr>
<th>Name</th>
<th>Comments</th>
<th>Reach</th>
<th>Wavelength/Source</th>
<th>Form-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>100GBase-SR10</td>
<td>10 lasers each operating at 10 Gb/s. Electrical I/O is 10 Gb/s.</td>
<td>100m (OM3)</td>
<td>850nm VCSEL Array</td>
<td>CFP, CXP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>125m (OM4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100GBase-SR4</td>
<td>4 Lasers each operating at 25Gb/s. Electrical I/O is retimed 25Gb/s</td>
<td>70m (OM3)</td>
<td>850nm VCSEL Array</td>
<td>CFP2, CFP4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100m (OM4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100GBase-LR4</td>
<td>4 lasers each operating at 25 Gb/s. Electrical I/O is retimed 25Gb/s</td>
<td>10km SMF</td>
<td>LAN WDM wavelengths</td>
<td>CFP, CFP2,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1295nm, 1300nm, 1305nm, 1310nm</td>
<td>CFP4</td>
</tr>
</tbody>
</table>

![CFG module diagram](image)

**Legend:**
- **RXMCLK:** Transmit Clock (default), Receive Clock (disabled)
- **TXMCLK:** Transmit Clock (default), Receive Clock (disabled)
- **RXDATA:** Receiver Data (default), Transmitter Data (disabled)
- **TXDATA:** Transmitter Data (default), Receiver Data (disabled)
- **REFCLK:** Synchronous with TXDATA

**Notes:**
- $M \neq N$
- 100GBASE-LR4
- 100GBASE-ER4
- 40GBASE-FR
- OC768 (40G, 43G)
- $M =$ Electrical lane number
- $N =$ Optical lane number
Major data center principle: parallelism

• Unlike smaller networks, data center networks exhibit massive parallelism
  – FatTrees have many many paths from source to destination
  – Computation spread across many processors located in many servers
  – A single storage service spread across 1000s of individual storage servers

• How to program such a beast?
  – MapReduce, Spark, GraphX, GraphLab, ...
Data Center Costs

• James Hamilton published basic 2008 breakdown
  - Servers: 45%
    - CPU, memory, disk
  - Infrastructure: 25%
    - UPS, cooling, power distribution
  - Power draw: 15%
    - Electrical utility costs
  - Network: 15%
    - Switches, links, transit
Data center power usage

- CPUs: 42.0%
- DRAM: 15.4%
- Disks: 14.3%
- Networking: 11.7%
- Misc.: 7.7%
- Power Overhead: 4.9%
- Cooling Overhead: 4.0%
Data center power efficiency

• What does power efficiency mean?
• How could you measure power efficiency?
  – For your own home
  – For a single computer
  – For a data center
• Can you directly compare
  – Facebook and Google?
  – Netflix.com and Hulu.com?
Quantifying energy-efficiency: PUE

- PUE = Power Usage Effectiveness
- Simply compares
  - Power used for computing
  - Total power used

- Historically cooling was a huge source of power
  - E.g., 1 watt of computing meant 1 Watt of cooling!

PUE = (Facility Power) / (Computing Equipment power)
Google’s “Chiller-less” Data Center

• Belgium
• Most of the year it is cool enough to not need cooling
• What about on hot days?
  – Shed load to other data centers!
What about “power saving” features on modern computers?

![Bar and Line Graph showing Performance to Power Ratio across various Target Loads and Average Power (W)]
Quantifying performance of a cluster

• Typically we think of performance in terms of the mean or median
  – Fine for a single processor/server
  – Not fine for an ensemble of 100s or 1000s of machines
  – Why?
Tail Tolerance: Partition/Aggregate

• Consider distributed memcached cluster
  – Single client issues request to S memcached servers
    • Waits until all S are returned
  – Service time of a memcached server is normal w/ \( \mu = 90\text{us}, \sigma = 50\text{us} \)
    • Based on microbenchmarks we took a few years ago
  – Service time of composed system is based on:

\[
\bar{\mathbf{v}} = \langle N(\mu, \sigma), N(\mu, \sigma), \ldots, N(\mu, \sigma) \rangle
\]
Partition/Aggregate simulation

Maximum Expected Latency (in us)

Simulated Number of Servers

99% N(90,50) distribution
50% N(90,50) distribution
Grounding the simulation

Empirically observed latency
99% latency single server
50% latency single server

Maximum Expected Latency (in us)

Latency (in us)

Simulated Number of Servers

Number of Servers
Tail tolerance: Dependent/Sequential pattern

• Consider iterative lookups in a service to build a web page
  – E.g., Facebook

• Issue request, get response, based on response, issue new request, etc…

• How many iterations can we issue within a deadline D?
  – For reference, Facebook limits # of RPCs to ~100
Dependent/Sequential pattern

• Service time of a web service is a function of the load
  – As is the variance

• M/G/1 Queuing model:
  – $\lambda$: rate of requests to a service (Poisson)
  – $\mu$: average service time
  – $\rho = \lambda/\mu$

\[
W = \frac{\rho + \lambda \mu \text{Var}(S)}{2(\mu - \lambda)}
\]

Via Pollaczek-Khinchine transformation
Dependent/Sequential pattern Simulation

Server load in requests/sec x(1000)

#Requests within SLA

- 50ms SLA stddev=2us
- 50ms SLA stddev=1us
Dependent/Sequential pattern
Validation