CSE 291: Data Center Networking

Spring 2015
Tu/Th 8:00-9:20am
George Porter
UC San Diego
Outline

- Course Mechanics
- Course Topics / Outline
- Introduction to data center networking
Audience

• Who should take this course
  – Those interested in learning about how large-scale Internet services are designed, built, deployed, and operated
    • With a specific focus on the network
  – Those wanting to explore the state-of-the-art in network design, especially design *at scale*
  – MS students with a concentration in Communication Networks
Specifics

- Professor: George Porter, gmporter@cs.ucsd.edu
  - Group office hours: 3104 CSE (typically Tue 9:30-10:30am, immediately after class)

- Teaching Staff
  - Yashar Asgarieh
  - Office Hours: TBD

- Course Web Page
  - http://www.cs.ucsd.edu/~gmporter/classes/sp15/cse291/

- Mailing list
  - Will be created based on the roster in the next few days
General learning objectives

• How to read a research paper in an objective manner.
• How to articulate your understanding and insights into a research paper.
• How to synthesize research themes and topics across multiple papers.
• How to design a research map, annotated bibliography, related work document, and a project proposal.
• Carry out a project related to data center networking.
• Presentation of research results.
Non-goals

• Listen to me talk for 75 minutes a day
• Passively “learn” the material
  – To really get the most from this class, you need to actively engage with the material!
  – In class discussion
  – Question assumptions, push back...
Grading

1. 15% In-class participation (the success of the class depends on your participation!)
2. 45% Paper summaries
3. 40% Course project, including a presentation at the end of the term (during the final exam period)
   • No exams (midterm or final)
Grading

1. 15% In-class participation (the success of the class depends on your participation!)
   – Simple—read the paper before class, and come to class with questions, thoughts, feedback, ...
Grading

2. 45% Paper summaries
   • This term we will be using HotCRP to manage paper summaries
     – Write a summary of the paper
     – Enumerate some strengths of the paper
     – Enumerate some weaknesses of the paper
     – Answer a couple of guided thought questions
   • Upload to hotcrp by the start of class
Grading

3. 40% Course project, including a presentation at the end of the term (during the final exam period)
   • Groups of 2-3
   • Propose project idea (I can help you come up with a project if needed)
   • Work on project during the term
     – Various resources available
     – OK to do a “paper” project if appropriate
   • Final presentation during final exam period
     – Instead of a final exam
     – All students need to be present! Don’t book flights/trips/etc. that cause you to leave before the final exam period.
Outline

• Course Mechanics
• Course Topics / Outline
• Introduction to data center networking
Course topics

• Topologies
• Measurements
• Transport and low-latency delivery
• The control plane
  – SDN

• Circuit switching vs. packet switching
• Bufferless interconnects
• Inter-data center transport
• Conclusion / project presentations

Open to swapping out papers on the syllabus for others that you are all more interested in.
(Reading list)
Data Center Fundamentals:
The Datacenter as a Computer

George Porter
CSE 291
March 31, 2015

*Includes material taken from Barroso et al., 2013, and UCSD 222a.
Much in our life is now on the web
The web is driven by data

Data + amazon.com = Product Recommendations

Data + Spotify = Custom Stations

Data + Google = Personalized Search
Data-driven, per-user customization

Data + amazon.com = Product Recommendations
Cloud Computing

• Elastic resources
  – Expand and contract resources
  – Pay-per-use
  – Infrastructure on demand

• Multi-tenancy
  – Multiple independent users
  – Security and resource isolation
  – Amortize the cost of the (shared) infrastructure

• Flexible service management
  – Resiliency: isolate failure of servers and storage
  – Workload movement: move work to other locations
Cloud Service Models

• **Software as a Service (Saas)**
  – Provider licenses applications to users as a service
  – e.g., customer relationship management, email, …
  – Avoid costs of installation, maintenance, patches, …

• **Platform as a Service (Paas)**
  – Provider offers software platform for building applications
  – e.g., Google’s App-Engine
  – Avoid worrying about scalability of platform

• **Infrastructure as a Service (Iaas)**
  – Provider offers raw computing, storage, and network
  – e.g., Amazon’s Elastic Computing Cloud (EC2)
  – Avoid buying servers and estimating resource needs
Data centers with 100,000+ servers

Microsoft

Google

Facebook
These things are **really** big

- Google: 100 billion searches per month
- Facebook: 1.15 billion users
- Amazon.com: 120+ million users
The need for rapid growth
The need for rapid growth


Web Created

The need for rapid growth


- Web Created
- Google’s 1st cluster (15 years)
The need for rapid growth


Web Created

Google’s 1st cluster (15 years)

(10 years)
Host Virtualization

- Multiple virtual machines on one physical machine
- Applications run unmodified as on real machine
- VM can migrate from one computer to another
VMM Virtual Switches
Building blocks of modern data centers
Top-of-Rack Architecture

• Rack of servers
  – Commodity servers
  – And top-of-rack switch

• Modular design
  – Preconfigured racks
  – Power, network, and storage cabling

• Aggregate to the next level
Racks of servers (Google)
Facebook
Google
Extreme Modularity

- Containers
- Many containers
Not just a collection of servers

• A data center isn’t just a “small internet”
• Why?
  – Administered as a single domain
  – Trusted administrators
  – No need to be compatible with the “outside world”
    • Except for traffic to/from users
  – No need for international standards bodies
    • Though why do standards help?
“Front-End” datacenter traffic

Data center

Wide-area Internet

Internet Users
• Data sizes driven by the content that users actually consume
  – Growth largely due to higher bitrate content (IP TV/movies, iPhone FaceTime)
• Mobile Internet source of new users
• Often constrained by the “last mile”
Multi-Tier Applications

- Applications consist of tasks
  - Many separate components
  - Running on different machines
- Commodity computers
  - Many general-purpose computers
  - Not one big mainframe
  - Easier scaling
“Back-end” datacenter traffic

- Back-end analytics:
  - Connections between information
  - “Users who bought X also bought Y”
- Key differentiator determining success
  - Facebook vs Friendster
  - Amazon vs Buy.com
- Large-scale “join” computations spanning thousands of nodes
  - Need bandwidth as well as all-to-all connectivity
- Sorting / Searching
- Collaborative Filtering
- Map/Reduce
- Distributed Key/Value stores
- Video storage, post-production, and transmission
Data-intensive application requirements

**All-to-all**
- Performance gated on latency of shuffle phases
  - Need high bisection bandwidth

**Gather/Scatter**
- Performance gated on speed of slowest RPC/parallel operation
  - Need low variance
Increasingly stringent network requirements

- Low one-way latency (10s of microseconds)
- 10/40 Gbps at TOR (and soon endhosts)
- Congestion-free operation/low queuing
- Dynamic traffic...
- ...and an increasingly dynamic topology
From networks to backplanes

• Before:
  – Network connects servers to users (FTP, telnet, …)
  – Massive computing = tightly coupled supercomputer
    • Proprietary interconnects
    • Working sets of data

• Today:
  – Servers connected to each other
  – Data-intensive, web-scale
  – Massive computing = Datacenters
    • Commodity
    • Datacenter network becomes the computing backplane
Data Center Challenges

• Traffic load balancing
• Support for VM migration
• Achieving bisection bandwidth
• Power savings / Cooling
• Network management (provisioning)
• Security (dealing with multiple tenants)
## Data Center Storage Example

### Photos @ Facebook

<table>
<thead>
<tr>
<th></th>
<th>April 2009</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>15 billion photos</td>
<td>65 billion photos</td>
</tr>
<tr>
<td></td>
<td>60 billion images</td>
<td>260 billion images</td>
</tr>
<tr>
<td></td>
<td>1.5 petabytes</td>
<td>20 petabytes</td>
</tr>
<tr>
<td><strong>Upload Rate</strong></td>
<td>220 million photos / week</td>
<td>1 billion photos / week</td>
</tr>
<tr>
<td></td>
<td>25 terabytes</td>
<td>60 terabytes</td>
</tr>
<tr>
<td><strong>Serving Rate</strong></td>
<td>550,000 images / sec</td>
<td>1 million images / sec</td>
</tr>
</tbody>
</table>

Finding a Needle in Haystack: Facebook's Photo Storage, OSDI’10
The storage hierarchy

One Server
- DRAM: 16 GB, 100 ns, 20 GB/s
- Disk: 2 TB, 10 ms, 200 MB/s
- Flash: 128 GB, 100 us, 1 GB/s

Local Rack (80 servers)
- DRAM: 1 TB, 300 us, 100 MB/s
- Disk: 160 TB, 11 ms, 100 MB/s
- Flash: 20 TB, 400 us, 100 MB/s

Cluster (30 racks)
- DRAM: 30 TB, 500 us, 10 MB/s
- Disk: 4.80 PB, 12 ms, 10 MB/s
- Flash: 600 TB, 600 us, 10 MB/s
Latency, bandwidth, and capacity

Latency (us) — dotted line
Bandwidth (MB/sec) — dashed line
Capacity (GB) — solid line

Local DRAM, Local Disk, Rack DRAM, Rack Disk, Datacenter DRAM, Datacenter Disk
Performance of flash

[Graph showing performance comparison between DRAM, Flash, and Disk across latency (us), Ops/sec, and $/GB.]
Traditional DC Topology

- Core
  - Layer-3 router

- Aggregation
  - Layer-2/3 switch

- Access
  - Layer-2 switch
  - Servers

- Internet

Data Center