COMP4611: Design and Analysis of Computer Architectures

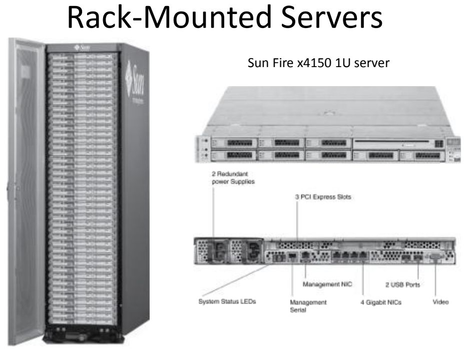
Warehouse-Scale Computers to Exploit Request-Level and Data-Level Parallelism

Lin Gu
CSE, HKUST

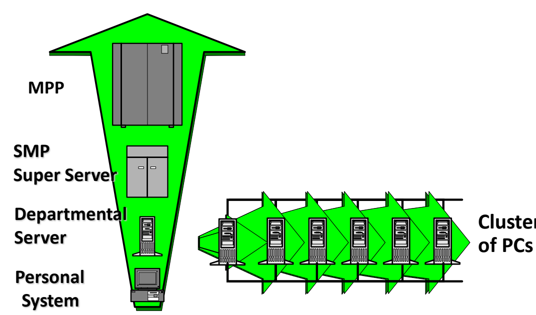
Server Computers

- Applications are increasingly run on servers
  - Web search, office apps, virtual worlds, …
- Requires large data center servers
  - Multiple processors, networks connections, massive storage
  - Space and power constraints
- Rack server equipment often in units of 1.75” (1U).
  - E.g., a 1U switch, a 2U server

Rack-Mounted Servers



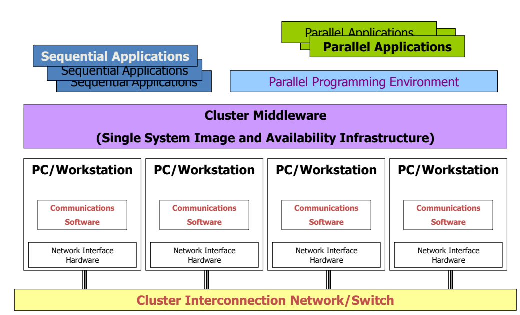
Scalability Vs. Cost



Motivations of using Clusters over Specialized Parallel Computers

- Individual PCs are becoming increasingly powerful
- Communication bandwidth between PCs is increasing and latency is decreasing (Gigabit Ethernet, Myrinet)
- PC clusters are easier to integrate into existing networks
- Typical low user utilization of PCs (<10%)
- Development tools for workstations and PCs are mature
- PC clusters are a cheap and readily available
- Clusters can be easily grown

Cluster Architecture

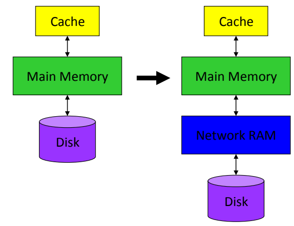


How Can we Benefit From Clusters?

- Given a certain user application
- **Phase 1**
  - If the application can be run fast enough on a single PC, there is no need to do anything else
  - Otherwise go to **Phase 2**
- **Phase 2**
  - Try to put the whole application on the DRAM to avoid going to the disk.
  - If that is not possible, use the DRAM of the other idle workstations
  - Network DRAM is 5 to 10 times faster than local disk

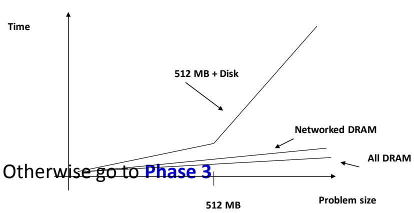
Remote Memory Paging

- Background
  - Application's working sets have increased dramatically
  - Applications require more memory than a single workstation can provide.
- Solution
  - Inserts the Network DRAM in the memory hierarchy between local memory and the disk
  - Swaps the page to remote memory



How Can we Benefit From Clusters?

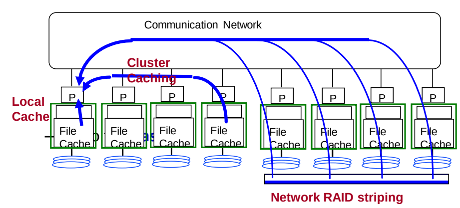
- In this case, the DRAM of the networked PCs behave like a ***huge cache*** system for the disk



- Otherwise go to **Phase 3**

How Can we Benefit From Clusters?

- **Phase 3**
  - If the network DRAM is not large enough, try using all the disks in the network in parallel for reading and writing data and program code (e.g., RAID) to speedup the I/O

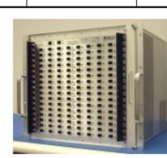


How Can we Benefit From Clusters?

- **Phase 4**
  - Execute the program on a multiple number of workstations (PCs) at the same time – **Parallel processing**
- **Tools**
  - There are many tools that do all these phases in a **transparent** way (except parallelizing the program) as well as load-balancing and scheduling.
    - Beowulf (CalTech and NASA) - USA
    - Condor - Wisconsin State University, USA
    - MPI (MPI Forum, MPICH is one of the popular implementations)
    - NOW (Network of Workstations) - Berkeley, USA
    - PVM - Oak Ridge National Lab./UTK/Emory, USA

What network should be used?

	Fast Ethernet	Gigabit Ethernet	Myrinet	10GbE
Latency	~120µs	~120 µs	~7 µs	10s of µs's
Bandwidth	~100Mbps peak	~1Gbps peak	~1.98Gbps real	10Gbps peak



2007 Top500 List

- Clusters are the fastest growing category of supercomputers in the TOP500 List.
  - 406 clusters (81%) in November 2007 list
  - 130 clusters (23%) in the June 2003 list
  - 80 clusters (16%) in the June 2002 list
  - 33 clusters (6.6%) in the June 2001 list
- 4% of the supercomputers in the November 2007 TOP500 list use Myrinet technology!
- 54% of the supercomputers in the November 2007 TOP500 list Gigabit Ethernet technology!

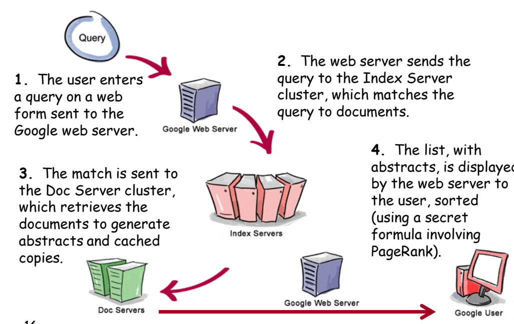
Introduction

- Warehouse-scale computer (WSC)
  - Provides Internet services
    - Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.
  - Differences with HPC "clusters":
    - Clusters have higher performance processors and network
    - Clusters emphasize thread-level parallelism, WSCs emphasize request-level parallelism
  - Differences with datacenters:
    - Datacenters consolidate different machines and software into one location
    - Datacenters emphasize virtual machines and hardware heterogeneity in order to serve varied customers

Introduction

- Important design factors for WSC:
  - Cost-performance
    - Small savings add up
  - Energy efficiency
    - Affects power distribution and cooling
    - Work per joule
  - Dependability via redundancy
  - Network I/O
  - Interactive and batch processing workloads
  - Ample computational parallelism is not important
    - Most jobs are totally independent
    - "Request-level parallelism"
  - Operational costs count
    - Power consumption is a primary, not secondary, constraint when designing system
  - Scale and its opportunities and problems
    - Can afford to build customized systems since WSC require volume purchase

Google



16

Google Requirements

- Google: search engine that scales at Internet growth rates
- Search engines: 24x7 availability
- Google : 600M queries/day, or AVERAGE of 7500 queries/s all day (old data)
- Google crawls WWW and puts up new index every 2 weeks (old data)
- Storage: 5.3 billion web pages, 950 million newsgroup messages, and 925 million images indexed, Millions of videos (very old data)
- Response time goal: < 0.5 s per search (old data)

17

Google

(Based on old data)

- Require high amounts of computation per request
- A single query on Google (on average)
  - reads hundreds of megabytes of data
  - consumes tens of billions of CPU cycles
- A peak request stream on Google
  - requires an infrastructure comparable in size to largest **supercomputer** installations
- Typical google Data center: 15000 PCs (Linux), 30000 disks: almost 3 petabyte!
- Google application affords easy parallelization
  - Different queries can run on different processors
  - A single query can use multiple processors
    - because the overall index is partitioned

18

Prgrm'g Models and Workloads

Batch processing framework: MapReduce

- **Map:** applies a programmer-supplied function to each logical input record
  - Runs on thousands of computers
  - Provides new set of key-value pairs as intermediate values
- **Reduce:** collapses values using another programmer-supplied function

Prgrm'g Models and Workloads

Example:

- **map (String key, String value)**:
  - **// key: document name**
  - **// value: document contents**
  - **for each word w in value**
    - **EmitIntermediate(w,"1"); // Produce list of all words**
- **reduce (String key, Iterator values):**
  - **// key: a word**
  - **// value: a list of counts**
  - **int result = 0;**
  - **for each v in values:**
    - **result += ParseInt(v); // get integer from key-value pair**
  - **Emit(AsString(result));**

Prgrm'g Models and Workloads

- **MapReduce runtime environment schedules map and reduce task to WSC nodes**
- **Availability:**
  - **Use replicas of data across different servers**
  - **Use relaxed consistency:**
    - **No need for all replicas to always agree**
- **Workload demands**
  - **Often vary considerably**

Computer Architecture of WSC

- **WSC often use a hierarchy of networks for interconnection**
- **Each rack holds dozens of servers connected to a rack switch**
- **Rack switches are uplinked to switch higher in hierarchy**
  - **Uplink has 48 / n times lower bandwidth, where n = # of uplink ports**
    - **"Oversubscription"**
  - **Goal is to maximize locality of communication relative to the rack**

Storage

- **Storage options:**
  - **Use disks inside the servers, or**
  - **Network attached storage through Infiniband**
  - **WSCs generally rely on local disks**
  - **Google File System (GFS) uses local disks and maintains at least three replicas**

Array Switch

- **Switch that connects an array of racks**
  - **Array switch should have 10 X the bisection bandwidth of rack switch**
  - **Cost of n-port switch grows as n^2**
  - **Often utilize content addressible memory chips and FPGAs**

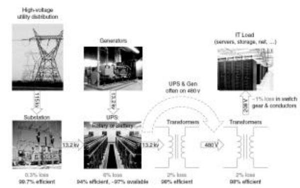
WSC Memory Hierarchy

Servers can access DRAM and disks on other servers using a NUMA-style interface

	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1,040	31,200
Disk capacity (GB)	2000	160,000	4,800,000

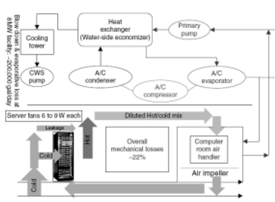
Infrastructure and Costs of WSC

- **Location of WSC**
  - **Proximity to Internet backbones, electricity cost, property tax rates, low risk from earthquakes, floods, and hurricanes**
- **Power distribution**



Infrastructure and Costs of WSC

- **Cooling**
  - **Air conditioning used to cool server room**
  - **64 F – 71 F**
    - **Keep temperature higher (closer to 71 F)**
  - **Cooling towers can also be used**
    - **Minimum temperature is "wet bulb temperature"**



Infrastructure and Costs of WSC

- **Cooling system also uses water (evaporation and spills)**
  - **E.g. 70,000 to 200,000 gallons per day for an 8 MW facility**
- **Power cost breakdown:**
  - **Chillers: 30-50% of the power used by the IT equipment**
  - **Air conditioning: 10-20% of the IT power, mostly due to fans**
- **How man servers can a WSC support?**
  - **Each server:**
    - **"Nameplate power rating" gives maximum power consumption**
    - **To get actual, measure power under actual workloads**
  - **Oversubscribe cumulative server power by 40%, but monitor power closely**

Measuring Efficiency of a WSC

- **Power Utilization Effectiveness (PEU)**
  - **= Total facility power / IT equipment power**
  - **Median PUE on 2006 study was 1.69**
- **Performance**
  - **Latency is important metric because it is seen by users**
  - **Bing study: users will use search less as response time increases**
  - **Service Level Objectives (SLOs)/Service Level Agreements (SLAs)**
    - **E.g. 99% of requests be below 100 ms**

Cost of a WSC

- **Capital expenditures (CAPEX)**
  - **Cost to build a WSC**
- **Operational expenditures (OPEX)**
  - **Cost to operate a WSC**

Cloud Computing

- **WSCs offer economies of scale that cannot be achieved with a datacenter:**
  - **5.7 times reduction in storage costs**
  - **7.1 times reduction in administrative costs**
  - **7.3 times reduction in networking costs**
  - **This has given rise to cloud services such as Amazon Web Services**
    - **"Utility Computing"**
    - **Based on using open source virtual machine and operating system software**