CS252
Graduate Computer Architecture

Lecture 10:
Network 3: Clusters, Examples

February 16, 2001

Prof. David A. Patterson

Computer Science 252

Spring 2001

Review: Networking

- Protocols allow hetereogeneous networking
  - Protocols allow operation in the presense of failures
  - Internetworking protocols used as LAN protocols
    => large overhead for LAN
- Integrated circuit revolutionizing networks as well as processors
  - Switch is a specialized computer
  - Faster networks and slow overheads violate of Amdahl's Law
- Wireless Networking offers new challenges in bandwidth, mobility, reliability, ...

Cluster

- LAN switches => high network bandwidth and scaling was available from off the shelf components
- 2001 Cluster = collection of independent computers using switched network to provide a common service
- Many mainframe applications run more "loosely coupled" machines than shared memory machines (next chapter/week)
  - databases, file servers, Web servers, simulations, and multiprogramming/batch processing
  - Often need to be highly available, requiring error tolerance and repairability
  - Often need to scale

Cluster Drawbacks

- Cost of administering a cluster of N machines ~ administering N independent machines
  vs. cost of administering a shared address space N processors multiprocessor ~ administering 1 big machine
- Clusters usually connected using I/O bus, whereas multiprocessors usually connected on memory bus
- Cluster of N machines has N independent memories and N copies of OS, but a shared address multiprocessor allows 1 program to use almost all memory
  - DRAM prices has made memory costs so low that this multiprocessor advantage is much less important in 2001

Cluster Advantages

- Error isolation: separate address space limits contamination of error
- Repair: Easier to replace a machine without bringing down the system than in an shared memory multiprocessor
- Scale: easier to expand the system without bringing down the application that runs on top of the cluster
- Cost: Large scale machine has low volume => fewer machines to spread development costs vs. leverage high volume off-the-shelf switches and computers
- Amazon, AOL, Google, Hotmail, Inktomi, WebTV, and Yahoo rely on clusters of PCs to provide services used by millions of people every day

Addressing Cluster Weaknesses

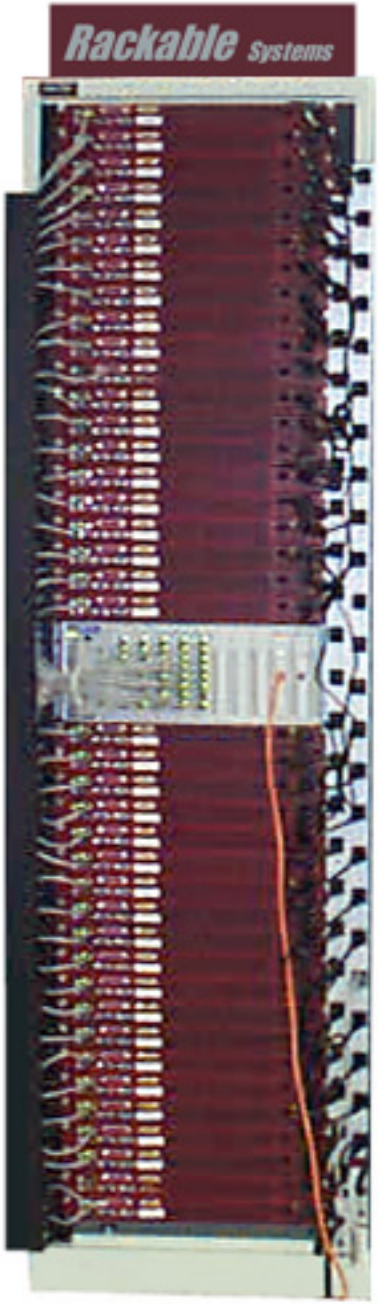
- Network performance: SAN, especially Inifiband, may tie cluster closer to memory
- Maintenance: separate of long term storage and computation
- Computation maintenance:
  - Clones of identical PCs
  - 3 steps: reboot, reinstall OS, recycle
  - At $1000/PC, cheaper to discard than to figure out what is wrong and repair it?
- Storage maintenance:
  - If separate storage servers or file servers, cluster is no worse?

Clusters and TPC Benchmarks

- "Shared Nothing" database (not memory, not disks) is a match to cluster
- 2/2001: Top 10 TPC performance 6/10 are clusters (4 / top 5)

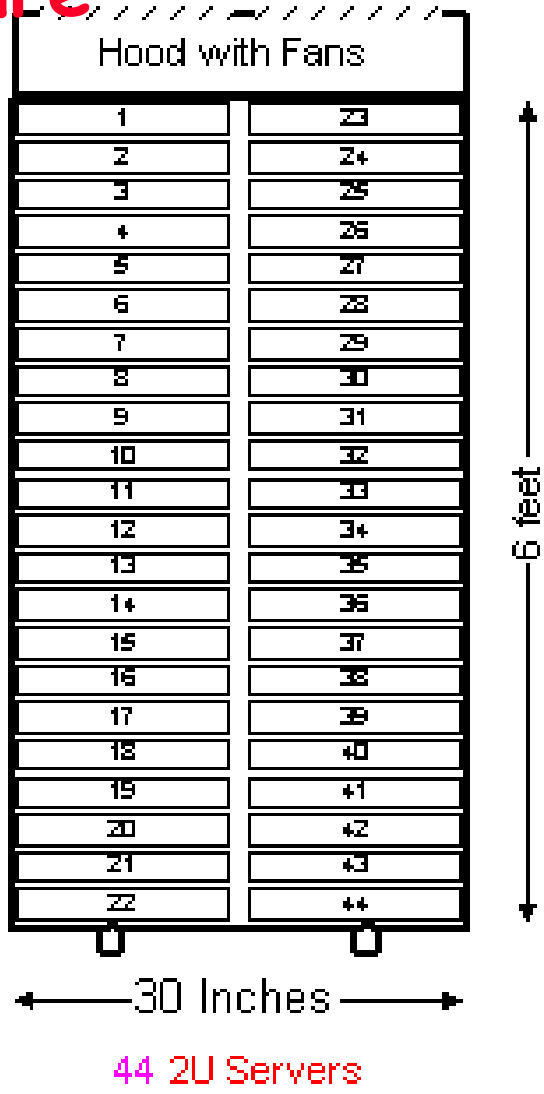
Putting it all together: Google

- Google: search engine that scales at growth Internet growth rates
- Search engines: 24x7 availability
- Google 12/2000: 70M queries per day, or AVERAGE of 800 queries/sec all day
- Response time goal: < 1/2 sec for search
- Google crawls WWW and puts up new index every 4 weeks
- Stores local copy of text of pages of WWW (snippet as well as cached copy of page)
- 3 collocation sites (2 CA + 1 Virginia)
- 6000 PCs, 12000 disks: almost 1 petabyte!



Hardware Infrastructure

- VME rack 19 in. wide, 6 feet tall, 30 inches deep
- Per side: 40 1 Rack Unit (RU) PCs +1 HP Ethernet switch (4 RU): Each blade can contain 8 100-Mbit/s EN or a single 1-Gbit Ethernet interface
- Front+back => 80 PCs + 2 EN switches/rack
- Each rack connects to 2 128 1-Gbit/s EN switches
- Dec 2000: 40 racks at most recent site



Google PCs

- 2 IDE drives, 256 MB of SDRAM, modest Intel microprocessor, a PC mother-board, 1 power supply and a few fans.
- Each PC runs the Linix operating system
- Buy over time, so upgrade components: populated between March and November 2000
  - microprocessors: 533 MHz Celeron to an 800 MHz Pentium III,
  - disks: capacity between 40 and 80 GB, speed 5400 to 7200 RPM
  - bus speed is either 100 or 133 MH
  - Cost: ~ $1300 to $1700 per PC
- PC operates at about 55 Watts
- Rack => 4500 Watts , 60 amps

Reliability

- For 6000 PCs, 12000s, 200 EN switches
- ~ 20 PCs will need to be rebooted/day
- ~ 2 PCs/day hardware failure, or 2%-3% / year
  - 5% due to problems with motherboard, power supply, and connectors
  - 30% DRAM: bits change + errors in transmission (100 MHz)
  - 30% Disks fail
  - 30% Disks go very slow (10%-3% expected BW)
- 200 EN switches, 2-3 fail in 2 years
- 6 Foundry switches: none failed, but 2-3 of 96 blades of switches have failed (16 blades/switch)
- Collocation site reliability:
  - 1 power failure,1 network outage per year per site
  - Bathtub for occupancy

CS 252 Administrivia

- Signup for meetings 12:00 to 2 Wed Feb 21
- Email project questionnaire Monday
- No lecture next Wednesday Feb 21

Google Performance: Serving

- How big is a page returned by Google?
  ~16KB
- Average bandwidth to serve searches

$$\frac{70{,}000{,}000/\text{day} \times 16{,}750\ \text{B} \times 8\ \text{bits/B}}{24 \times 60 \times 60}$$

$$= 9{,}378{,}880\ \text{Mbits}/86{,}400\ \text{secs}$$

$$= 108\ \text{Mbit/s}$$

Google Performance: Crawling

- How big is a text of a WWW page? ~4000B
- 1 Billion pages searched
- Assume 7 days to crawl
- Average bandwidth to crawl

$$\frac{1{,}000{,}000{,}000 \text{ pages} \times 4000 \text{ B} \times 8 \text{ bits/B}}{24 \times 60 \times 60 \times 7}$$

$$= 32{,}000{,}000 \text{ Mbits} / 604{,}800 \text{ secs}$$

$$= 59 \text{ Mbit/s}$$

Google Performance: Replicating Index

- How big is Google index? ~5 TB
- Assume 7 days to replicate to 2 sites, implies BW to send + BW to receive
- Average bandwidth to replicate new index

$$\frac{2 \times 2 \times 5{,}000{,}000 \text{ MB} \times 8 \text{ bits/B}}{24 \times 60 \times 60 \times 7}$$

$$= 160{,}000{,}000 \text{ Mbits} / 604{,}800 \text{ secs}$$

$$= 260 \text{ Mbit/s}$$

Colocation Sites

- Allow scalable space, power, cooling and network bandwidth plus provide physical security
- charge about $500 to $750 per Mbit/sec/month
  - if your continuous use measures 1- 2 Gbits/second

to $1500 to $2000 per Mbit/sec/month

  - if your continuous use measures 1-10 Mbits/second
- Rack space: costs $800 -$1200/month, and drops by 20% if > 75 to 100 racks (1 20 amp circuit)
  - Each additional 20 amp circuit per rack costs another $200 to $400 per month
- PG&E: 12 megawatts of power, 100,000 sq. ft./building, 10 sq. ft./rack => 1000 watts/rack

Google Performance: Total

- Serving pages: 108 Mbit/sec/month
- Crawling: 59 Mbit/sec/week, 15 Mbit/s/month
- Replicating: 260 Mbit/sec/week, 65 Mb/s/month
- Total: roughly 200 Mbit/sec/month
- Google's Collocation sites have OC48 (2488 Mbit/sec) link to Internet
- Bandwidth cost per month? ~$150,000 to $200,000
- 1/2 BW grows at 20%/month

Google Costs

- Collocation costs: 40 racks @ $1000 per month + $500 per month for extra circuits

= ~$60,000 per site, * 3 sites

~$180,000 for space

- Machine costs:
- Rack = $2k + 80 * $1500/pc + 2 * $1500/EN
  = ~$125k
- 40 racks + 2 Foundry switches @$100,000
  = ~$5M
- 3 sites = $15M
- Cost today is $10,000 to $15,000 per TB

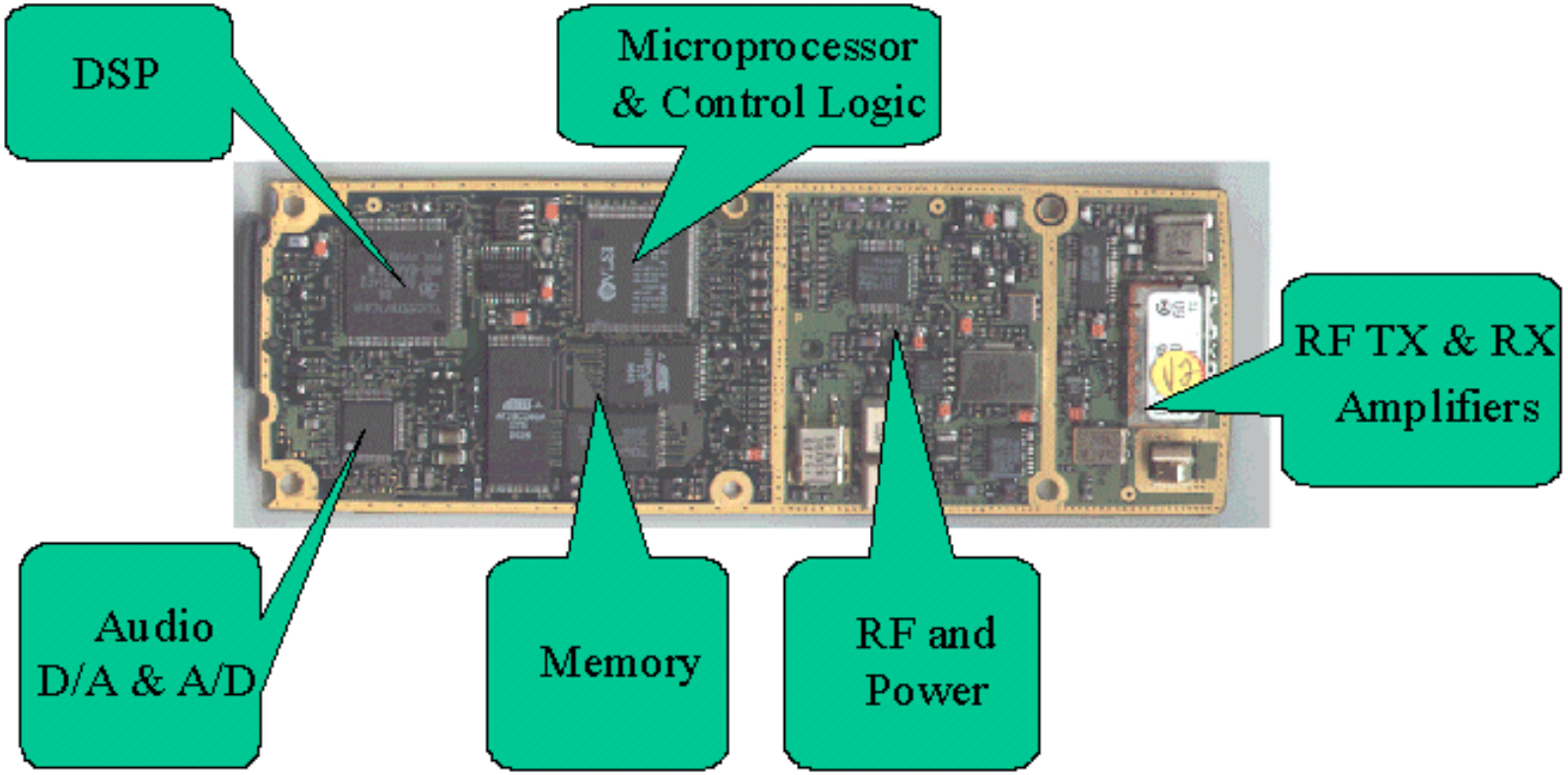
Comparing Storage Costs: 1/2001

- Google site, including 3200 processors and 0.8 TB of DRAM, 500 TB (40 racks) $10k - $15k/ TB
- Compaq Cluster with 192 processors, 0.2 TB of DRAM, 45 TB of SCSI Disks (17+ racks) $115k/TB (TPC-C)
- HP 9000 Superdome: 48 processors, 0.25 TB DRAM, 19 TB of SCSI disk = (23+ racks) $360k/TB (TPC-C)

Putting It All Together: Cell Phones

- 1999 280M handsets sold; 2001 500M
- Radio steps/components: Receive/transmit
  - Antenna
  - Amplifier
  - Mixer
  - Filter
  - Demodulator
  - Decoder

Putting It All Together: Cell Phones



- about 10 chips in 2000, which should shrink, but likely separate MPU and DSP
- Emphasis on energy efficiency

From "How Stuff Works" on cell phones: www.howstuffworks.com

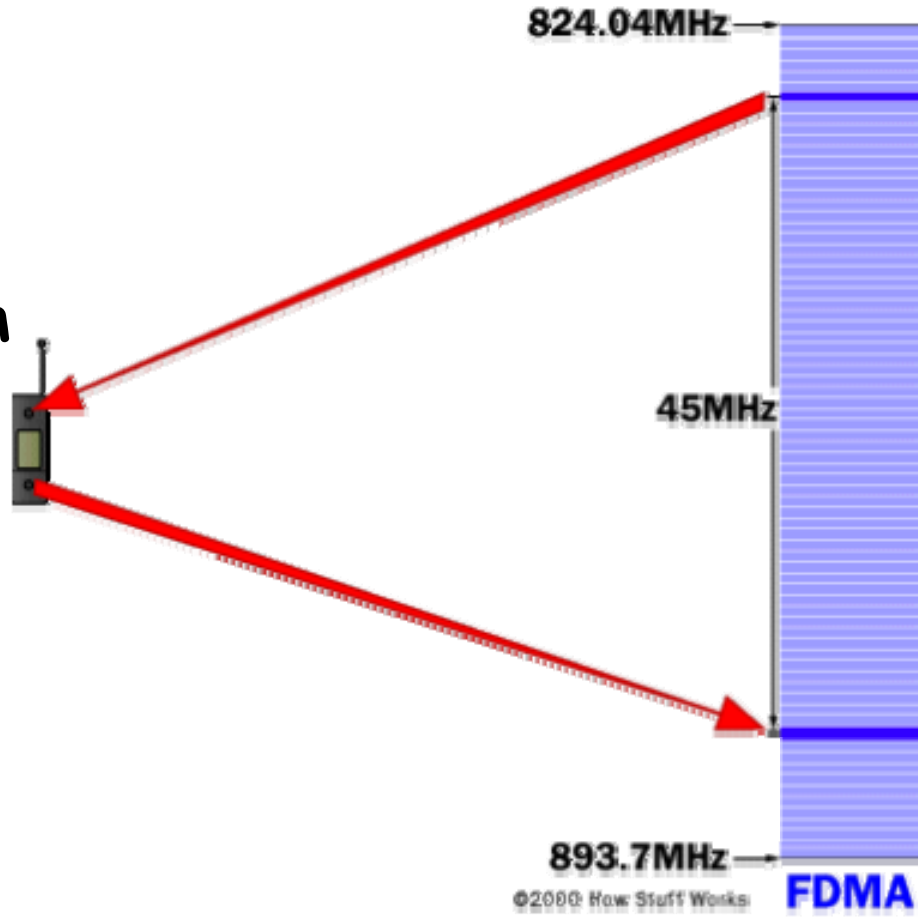
Cell phone steps (protocol)

- Find a cell
  - Scans full BW to find stronger signal every 7 secs

1. Local switching office registers call
   - records phone number, cell phone serial number, assigns channel
   - sends special tone to phone, which cell acks if correct
   - Cell times out after 5 sec if doesn't get supervisory tone
- Communicate at 9600 b/s digitally (modem)
  - Old style: message repeated 5 times
  - AMPS had 2 power levels depending on distance (0.6W and 3W)

Frequency Division Multiple Access (FDMA)

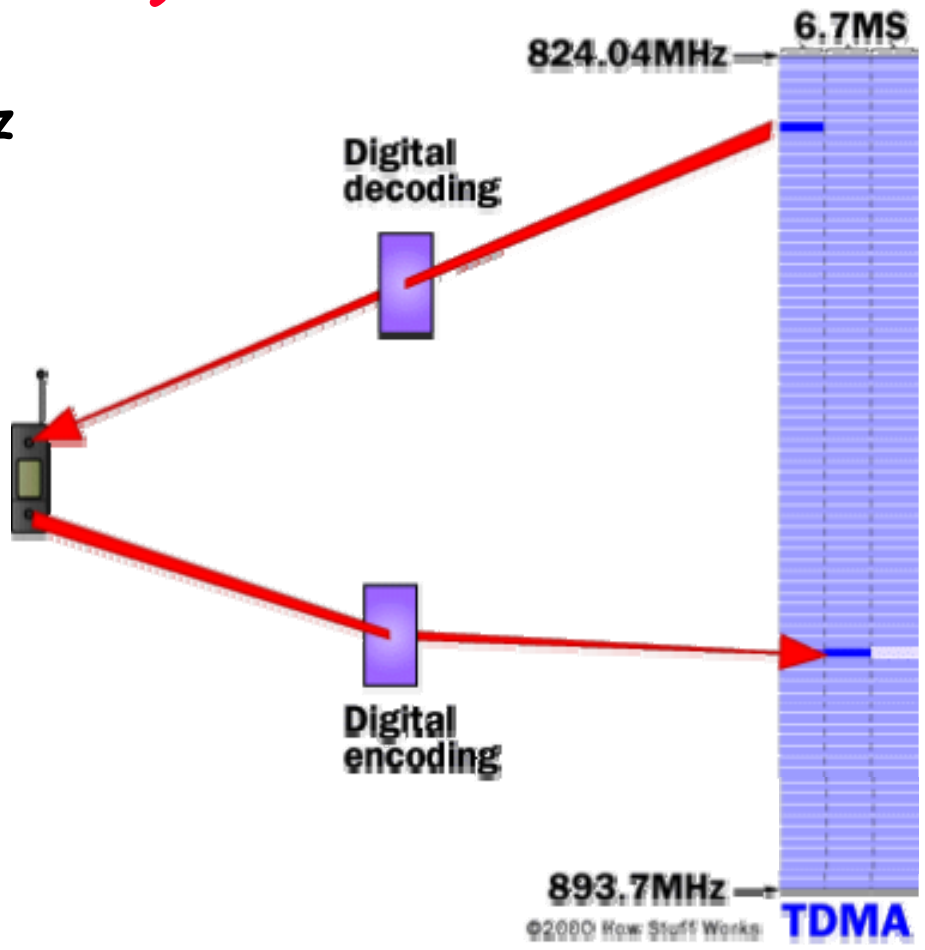
- FDMA separates the spectrum into distinct voice channels by splitting it into uniform chunks of bandwidth
- !st generation analog



From "How Stuff Works" on cell phones: www.howstuffworks.com

Time Division Multiple Access (TDMA)

- a narrow band that is 30 kHz wide and 6.7 ms long is split time-wise into 3 time slots.
- Each conversation gets the radio for 1/3 of time.
- Possible because voice data converted to digital information is compressed so
- Therefore, TDMA has 3 times capacity of analog
- GSM implements TDMA in a somewhat different and incompatible way from US (IS-136); also encrypts the call

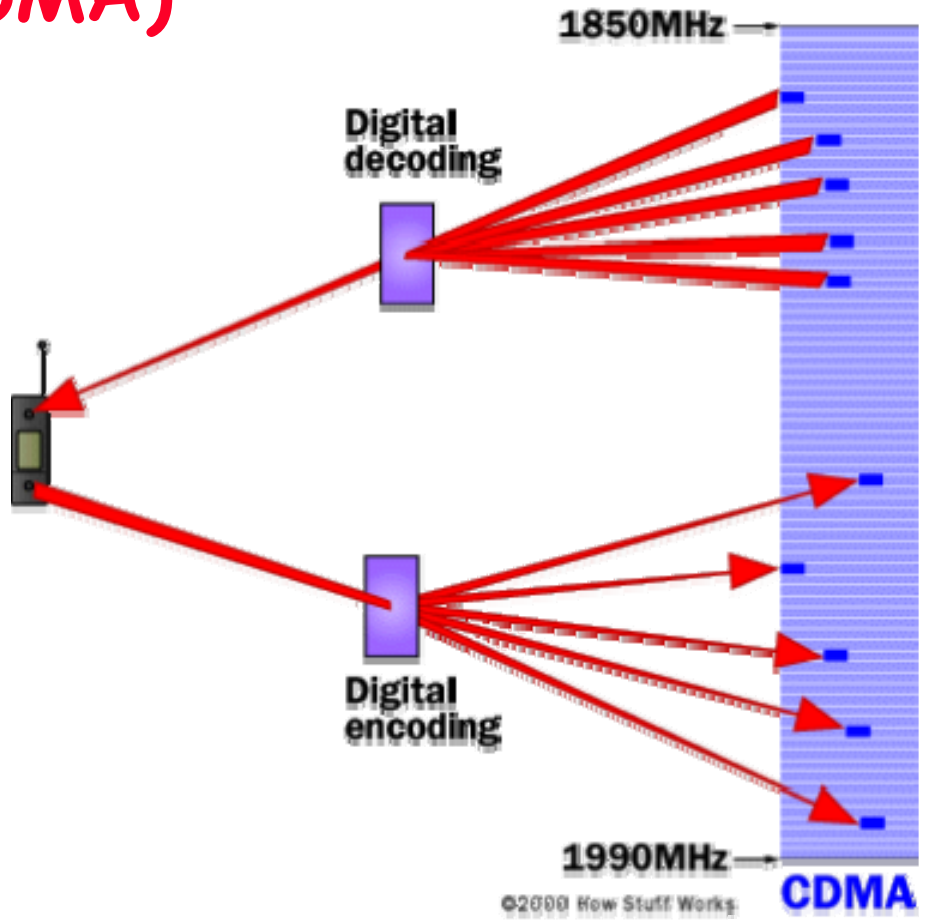


From "How Stuff Works" on cell phones: www.howstuffworks.com

Code Division Multiple Access (CDMA)

- CDMA, after digitizing data, spreads it out over the entire bandwidth it has available.
- Multiple calls are overlaid over each other on the channel, with each assigned a unique sequence code.
- CDMA is a form of spread spectrum; All the users transmit in the same wide-band chunk of spectrum.
- Each user's signal is spread over the entire bandwidth by a unique spreading code. same unique code is used to recover the signal.
- GPS for time stamp

. Between 8 and 10 separate calls space as 1 analog call



From "How Stuff Works" on cell phones: www.howstuffworks.com

Cell Phone Towers

From "How Stuff Works" on cell phones: www.howstuffworks.com

If time permits

- Discuss Hennessy paper. "The future of systems research." Computer, vol.32, (no.8), IEEE Comput. Soc, Aug. 1999
- Microprocessor Performance via ILP Analogy?
- What is key metric if services via servers is killer app?
- What is new focus for PostPC Era?
- How does he define availability vs. textbook definition?

Amdahl's Law Paper

- What was Amdahl's Observation?
- What is Amdahl's Law?