And NOW for Something Completely Different

Dave Patterson

Computer Science Division

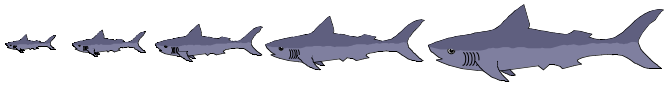
University of California, Berkeley

http://www.cs.berkeley.edu/~patterson/

Outline

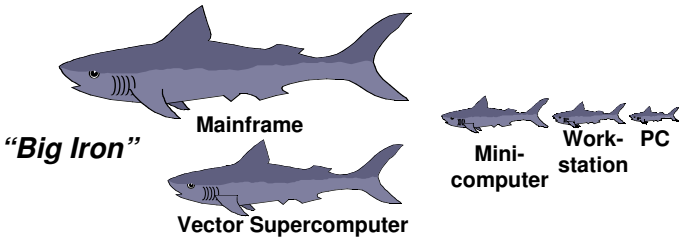
- Taxonomy and State of Computer Industry
- Timeliness of Networks of Workstations (NOW)
- Services on NOWs vs. Conventional Computers
- Research Challenges to Enable NOWs
- Conclusions

Biological Food Chain Picture



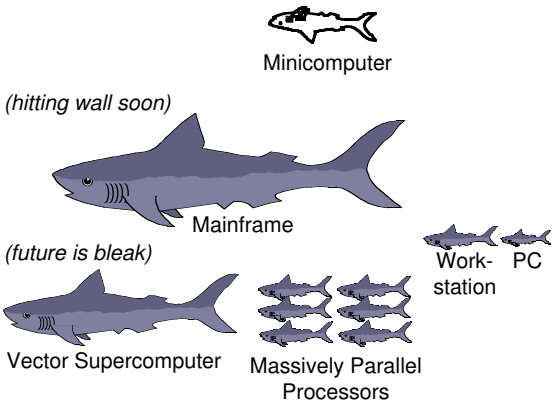
NOW 3

1985 Computer Food Chain

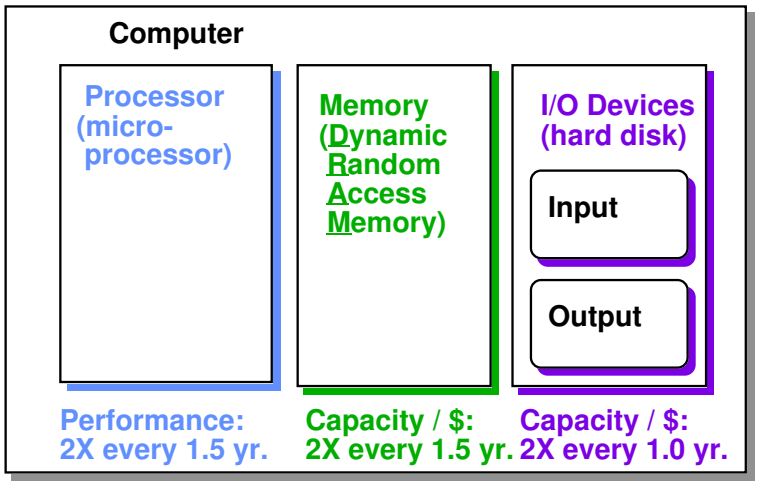


NOW 4

1995 Computer Food Chain

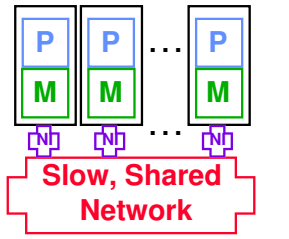


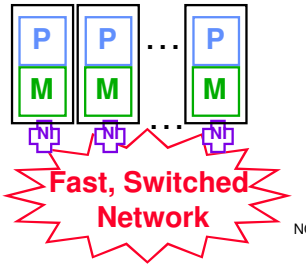
Classic Components of a Computer



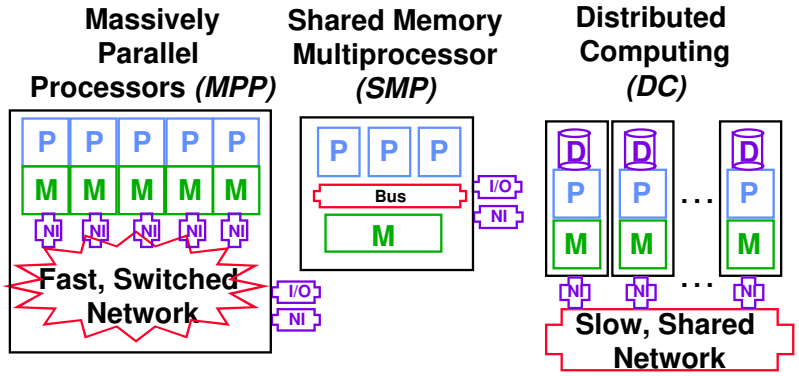
Networks to Connect Computers

- Shared Medium ("bus")
  - Cheap, easy to expand
  - Single conversation at a time
  - Often in Local Area Network(LAN)
  - e.g., AppleTalk, Ethernet
- Switched Medium ("switch")
  - Fast, leveraging semiconductors
  - Multiple conversations at a time
  - Often in Wide Area Network(WAN)
  - e.g., ATM

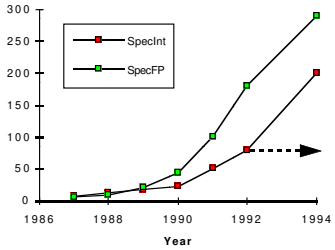




Multiple Computer Taxonomy



Engineering time costs performance



- When performance improves at 50% per year
  - 1 year lag => 2/3 of desktop performance
  - 2 year lag => 1/2 of desktop performance
- Massively Parallel Procs typically 1.5-3 yr. lag
- Shared Memory Procs often 0.5-1 year lag

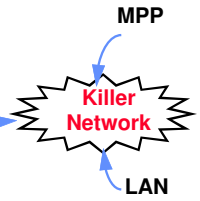
Dis-Economy of scale for Big Iron

- Small (cheap) systems have large volume
- Learning curve: *"When volume doubles, costs reduce by 10%"**
  - at 1,000x volume => costs differ by 3x
  - Big Iron can have higher cost even if components were the same
- Extra engineering cost is amortized over fewer machines
  - same for sales, maintenance, . . .
  - hence Big Iron can have higher price even if cost was the same

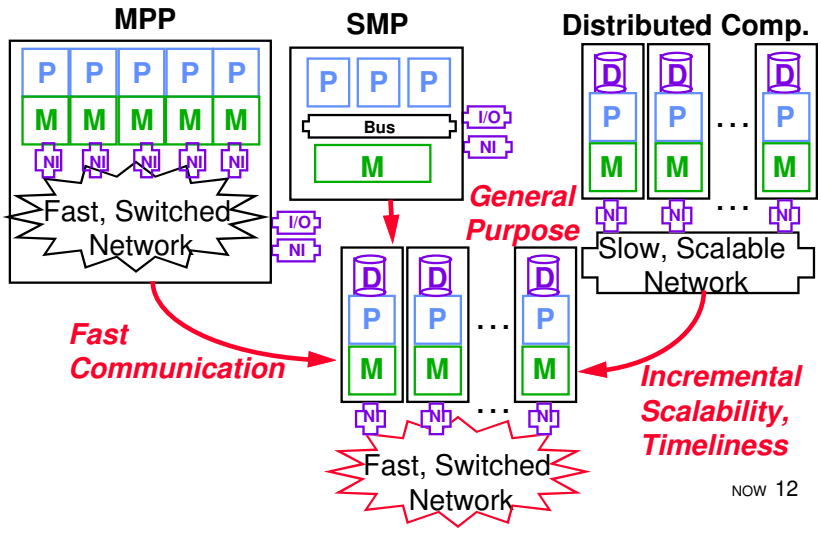
** A DEC View of Computer Engineering* by C. G. Bell, J. C. Mudge, and J. E. McNamara, Digital Press, Bedford, MA., 1978

Goal for Clusters or Networks of PCs/ Workstations (NOW)

- **Impact the design of Large Scale Systems**
  - solve large problems
  - serve large number of users
  - computing infrastructure of building, department, laboratory, ship, . . .
- **The Basis for building large-scale systems is:**
- **Complete mass-produced platforms**
  - "killer" micro, memory system, disk, O/S,
  - plus SOFTWARE
  - workstations or PCs
- **Scalable (switched) Networks**
  - high bandwidth, low latency, low error rate
  - ATM, Switched Ethernet, repackaged MPP: whoever wins



Cluster/Network of Workstations (NOW)



Why NOW now?

- **Building block is big enough** (v. PC AT)
  - 1995 workstation CPU ~ 1/3 supercomputer CPU
  - heavily under-utilized
- **Networks are faster**
  - Higher link bandwidth (v. 10 Mbit Ethernet)
  - Switch based networks coming (ATM)
  - New simple & fast paradigm for net. interfaces (Active Msgs)
- **Input/Output bottleneck**
  - Use network as I/O system interconnect (software RAID)
- **Demise of mainframes, supercomputers & MPPs**
  - "Big Iron" design => delays, poor cost/performance
- **Enables unique system design wins**
  - remote memory is much "closer" than local disk

Outline

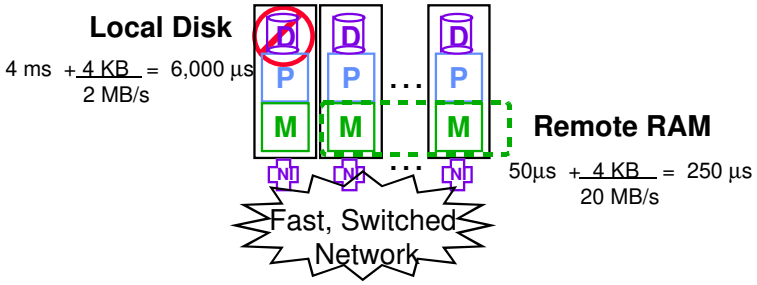
- Taxonomy and State of Computer Industry
- Timeliness of Networks of Workstations (NOW)
- **Services on NOWs vs. Conventional Computers**
- **Research Challenges to Enable NOWs**
- **Conclusions**

NOW: New Conceptual Framework

- HW: Fast Changing Technology => must reduce delays => connect off-the-shelf HW & SW
- SW: Offer Global Services on Local Devices

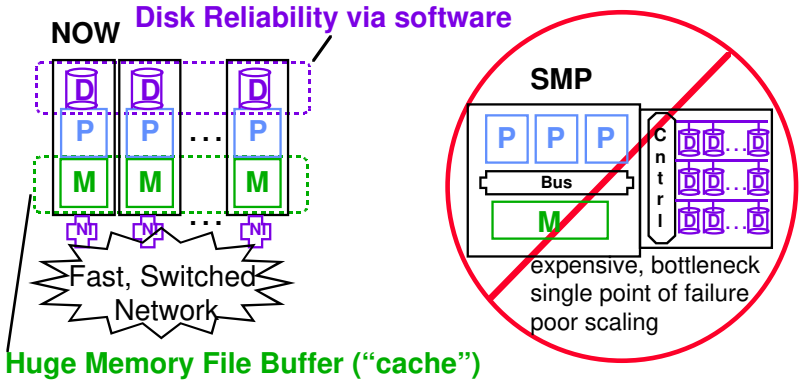
Problem	Big Iron Solution	NOW Solution
Virtual Memory	Local Disk	Remote RAM
File Service	Shared Mem. Proc.	Serverless File Sys.
Mass Storage	RAID/Tape Library	Massive Disk Array
High Speed Net	Gigabit Network	Striped Network
Number Crunch	MPP, Vector SMP	NOW
WWW Server	Fault Tolerant SMP	NOW

NOW Example: Fast Virtual Memory ("Network RAM")



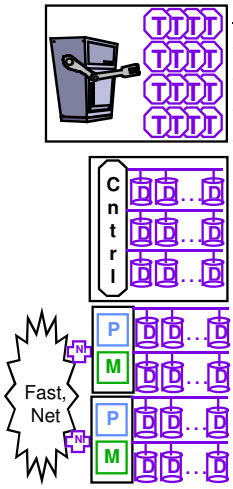
- Current Virtual Memory systems page to local disk => milliseconds
- Instead, go over network to Remote RAM => microseconds

NOW Example: Truly Distributed F.S. ("Serverless File Systems")



Huge Memory File Buffer ("cache")
- **private and global**
- **dynamic location (fault tolerance)**
- **latency critical**

NOW Example: Massive Storage ("Massive Array of Inexpensive Disks")



Tape Library: Slow, Improving slowly
- **Access time => secs, reading 10 TB => 1 week**
- **Tape MB/$ improving 1.4X / year**
- **Disks improving MB/$ by 2.0X / year since 1991**
- **MB/$ cross over 1 disk vs. tape library ≈ 2003?**

RAID: Not timely, Expensive to scale
- **RAID systems take ≈ 1 year to ship system based on new disk ("Big Iron")**
- **RAID systems cost ≈ 5X - 10X MB/$ as a single PC disk (higher margin)**

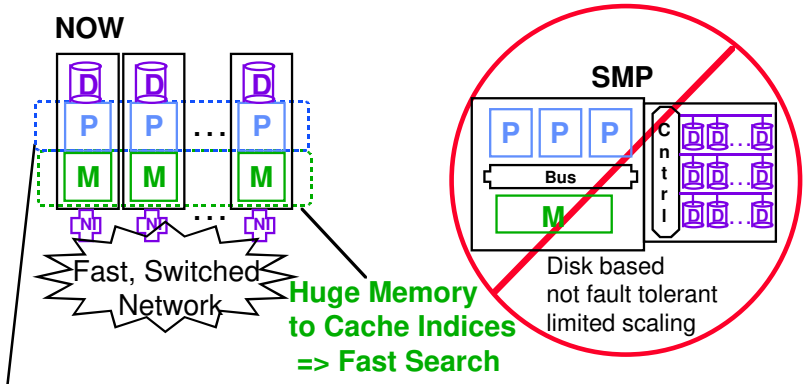
"MAID": Inexpensive Scaling
- **off-the-shelf racks, cables, computers, NIs, switches + SW RAID + maximum disks/computer**
- **Customer installs "just in time" disks, gets volume discount (≈25% @ 100)**

MB/$, Reliability 1000 disks ≈ 1 disk?

Apps in a Brave, New World (WWW)

- **Before: Copy of Software at each site**
  - Run software locally
  - Tell user community when unavailable when due to maintenance
  - Software running many places, so many users continue during down times of any one site
  - Each site has HW to run SW=> more users, more HW
- **After: Single Internet Web Server**
  - Run Software at ONE site on Internet
  - Downtime due to HW, OS, SW crash or upgrades looks like unreliable Web service => use a different service?
  - More users "overnight" depending on popularity, so single site hardware needs to scale incrementally, quickly, no upper limit
- **NOW is ideal Web server delivery vehicle?**

NOW Example: WWW Search Engine (Scalable, Available Searcher)



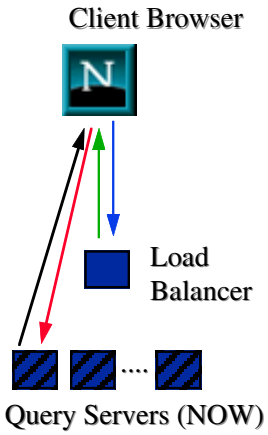
Scale processing & memory as workload increases

Node Failure => some searches go to disk, still available

Graceful Software upgrade for 365x24 hr. availability

Inktomi: Full-Text Index of the WWW

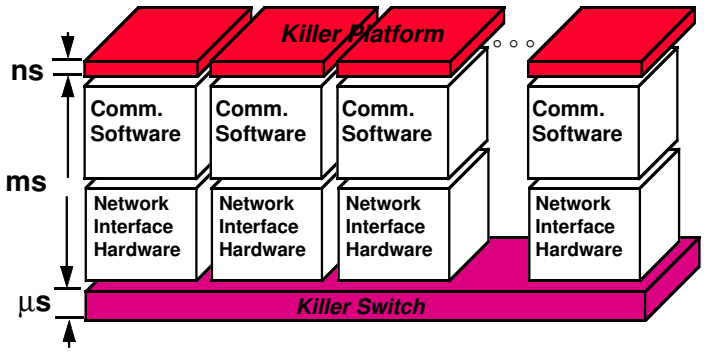
- **Keyword queries on 1.3 Million Documents (+ 5M URLs)**
- **Constructed on NOW: 4 SPARCstation 20s with fast communication software ("Active Messages") over fast switched network ("Myrinet")**
- **30X speed of commercial system ("Lycos") with fewer machines**
- **http://inktomi.berkeley.edu/**
  - **"inktomi" = name of mythical spider of Plains Indians; also ("ink-to-me")**
  - **Developed by Prof. Eric Brewer & Paul Gauthier**
  - **publicly available today**



Outline

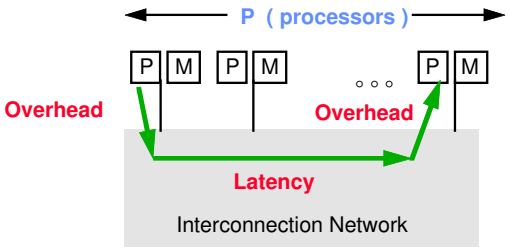
- Taxonomy and State of Computer Industry
- Timeliness of Networks of Workstations (NOW)
- Services on NOWs vs. Conventional Computers
- **Research Challenges to Enable NOWs**
- **Conclusions**

Challenge: realizing fast communication



The "last two inches" of network

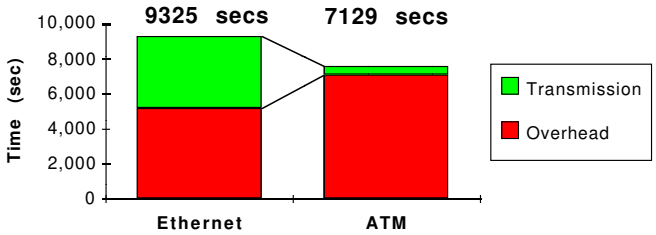
Communication Cost



- Network Latency incurred in sending message between nodes (1-way)
- Network Bandwidth ≈ 1/ Latency : Peak MB/s
- Processor Overhead to send *or receive* a message

Importance of Overhead

- Network File System (NFS) trace over 1 week: 95% msgs < 200 bytes
- Shared medium network (Ethernet): 9 Megabits/second BW, 456 μsecs overhead
- Switched (ATM): 78 Mbit/s BW, 626 μsecs ovhd.



- Bandwidth misleading metric?
  (625 μsec overhead ATM vs. 155 Mb/s BW ATM?)

NOW Research Agenda

Narrow Agenda

- Fast Communication

Broad Agenda (reinventing services/ WWW apps)

- Serverless file systems
- Network RAM
- WWW Search Engine/WWW history

Long Term Agenda

- System Management
- Heterogeneity (HW & OS)
- Hot Upgrades (HW & OS)
- Global Resource Allocation
- How to Build Fault Tolerant Applications
- Security

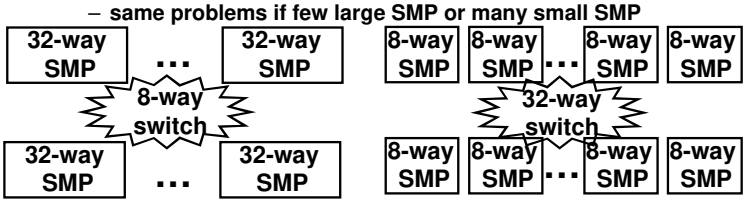
...

Conclusion

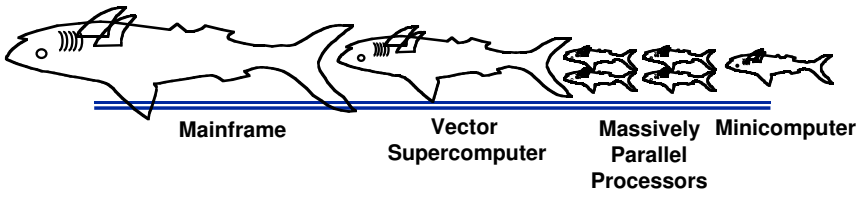
- Future: Lots of people, Lots of computers
  - some on desk, some centralized, some over WWW
- Networks of Workstations/PCs offer attractive solutions for large scale problems: scaling, fault isolation & tolerance, cost/performance
- Challenges for NOW: fast communication, global services on local devices, system management, hot upgrades, security
- Web servers may simplify NOW challenges

Computer Science Conclusion

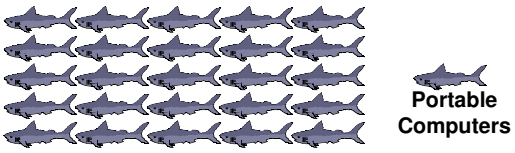
- Future: Lots of people, Lots of computers
  - some on desk, some centralized, some over Internet
- Overwhelming desire for all computers at site to be connected logically as a single system
  - same problems if few large SMP or many small SMP



- Prediction: As problems inside the box (HW &SW) become more of a solved problem, the new CompSci research challenges (HW &SW) will be connecting boxes together



2005 Computer Food Chain



To Study Further

Anderson,Culler, Patterson "A Case for Networks of Workstations: NOW"*IEEE Micro*, Feb. 1995.

Baskett, Hennessy "Microprocessors: from desktops to supercomputers," *Science*, 13 Aug. 1993, v 261, (5123): 864-871.

Kronenberg, Levy, Strecker "Vax Clusters: A closely-coupled distributed system," *ACM Transactions on Computer Systems*, May 1986 v 4(3):130-146.

Sep. 1995 *Scientific American*, Information Technologies.

Pfister, *In Search of Clusters – the coming battle in lowly parallel computing*, Prentice Hall, 1995, 415 pages.

World Wide Web: http://now.cs.berkeley.edu/

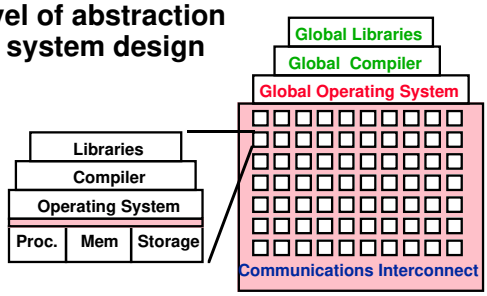
Backup Slides

(The following slides are only used for answering questions)

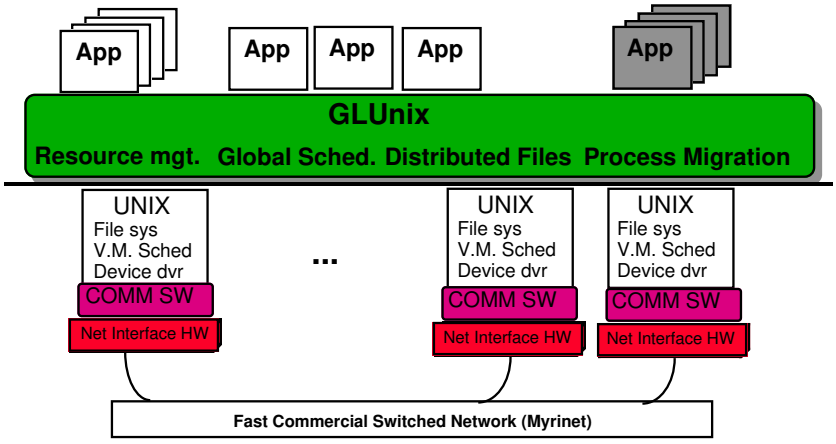
Technological Motivation

- **Convergence of LAN, MPP, and Telecommunications Networking**
- **Dominance of the complete desktop computer system**

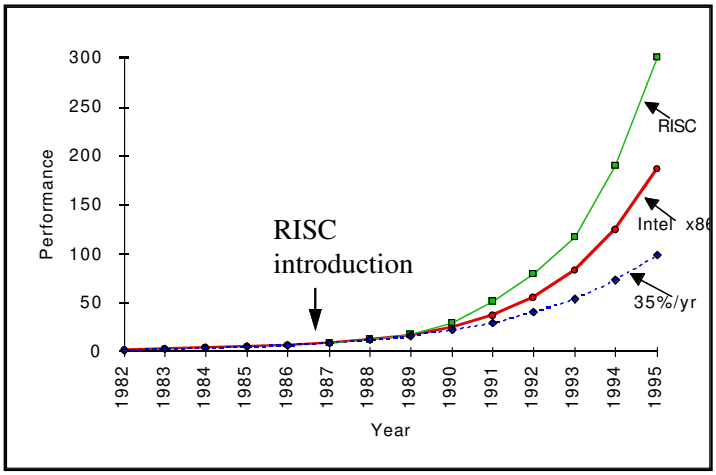
=> Raise the level of abstraction in large-scale system design



Berkeley NOW System Architecture

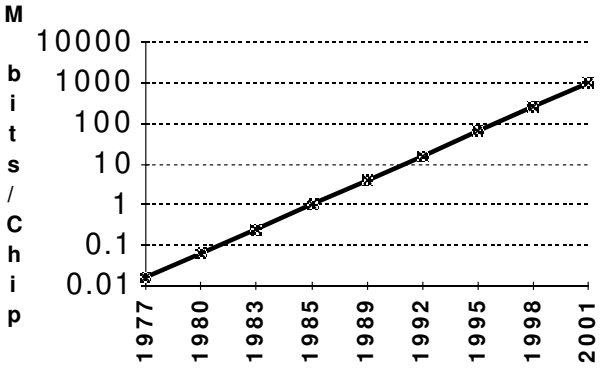


Processor Performance: 2X / 1.5 Yr.



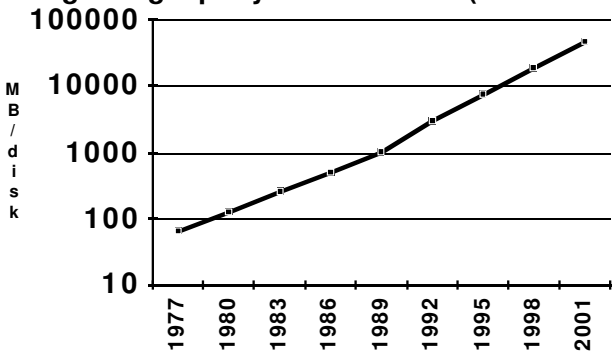
DRAM capacity growth

- **Trend: 4X every 3 years:**

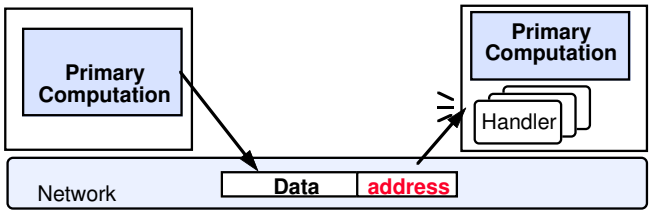


Magnetic Disks

- **Trend: Capacity & $/MB: Improving 27% / year (50X since 1977); since 1990 at 60% / year**
- **Improving areal density (bits/sq.in.) vs. a single target per year like DRAM (3 GB for 92)**



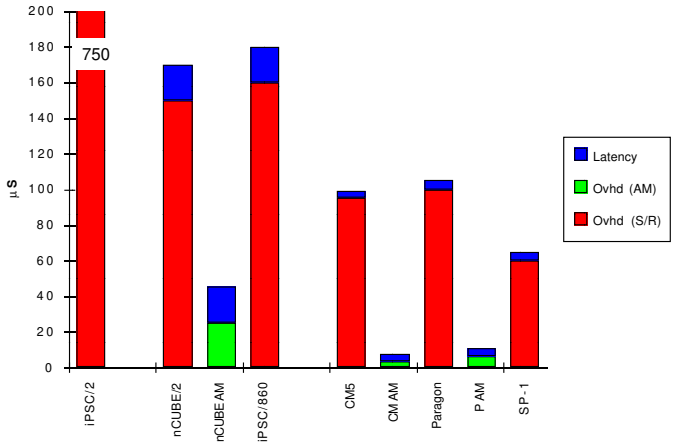
Active Messages



- **User generates msg => user knows how to handle it**
  - **Handler executes directly upon arrival**
  - **pulls message out of the network layer and integrates it into the computation or replies**
- **Simple protection, scheduling, allocation, liveness, . . .**

von Eicken, Culler, Goldstein, Schauser. "Active Messages: A Mechanism for Integrated Communication and Computation", ISCA 1992.

Active Messages on MPPs



Fast Trusted Area Communication

=> Translate MPP experience to NOW:

Active Messages plus

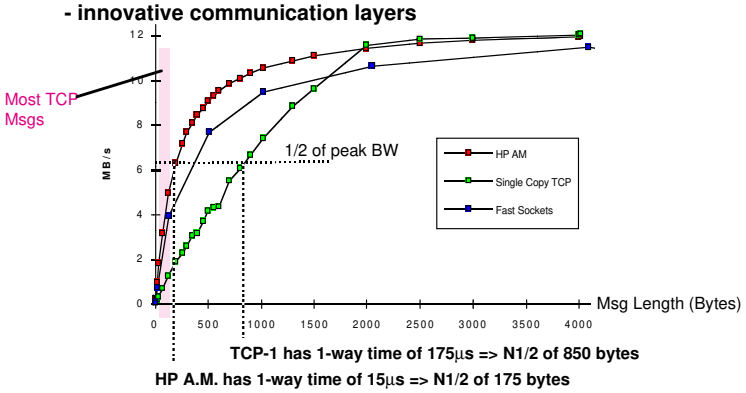
- **General Purpose Environment**
- **Serious Multiprogramming**
  - **Protection, naming, Resource Sharing**
- **Full Function Operating System**
  - **page faults, daemons, . . .**
- **Packet Loss**
  - **infrequent (if we do our job right)**

Traditional LAN: ~ 1,000 μs 1-way

NOW Goal: 10 μs 1-way

Prototype Comm. Performance

4 HP 9000/735 99MHz running HP UX
- experimental Medusa FDDI net. int. on the graphics bus
- innovative communication layers



TCP-1 has 1-way time of 175μs => N1/2 of 850 bytes

HP A.M. has 1-way time of 15μs => N1/2 of 175 bytes

Martin. "HPAM: An Active Message Layer for Networks of HP Workstations", Hot Interconnects 1994.

Why it works

- Careful engineering of a simple abstraction
- Closer binding between network and user
  - external scheduler handles resource sharing
- Stateful NI Card + Request/Response permits very simple approach to
  - flow-control
  - time-out and retry
- Build traditional abstractions on top
  - MPI, Sockets, . . .

XFS Performance

Time to transfer 8KB block:

	Remote Mem	Remote Disk
Ethernet	6.9 ms	21.9 ms
ATM (155Mb/s)	1.1 ms	16.1 ms

Simulation results: (Sprite file trace)

	Miss Rate	Response Time
Client/Server	16%	2.8 ms
Coop Caching	8%	1.6 ms

(42 WS, 16 MB/WS, 128 MB/server)

Dahlin, Wang, Anderson, and Patterson, “Cooperative Caching: Using Remote Client Memory to Improve File System Performance”, OSDI 1994.

Why it works

- **Disk access is the slowest part of the system**
- **Small increase in local miss rate**
  - **=> More communication (but with less contention)**
- **Larger decrease in global miss rate**
  - **=> Fewer disk accesses**

- **Selective (almost greedy) replacement**
  - **inactive workstations become more "server-like"**
  - **active workstations hold onto what they use**

Performance of Global Climate Model

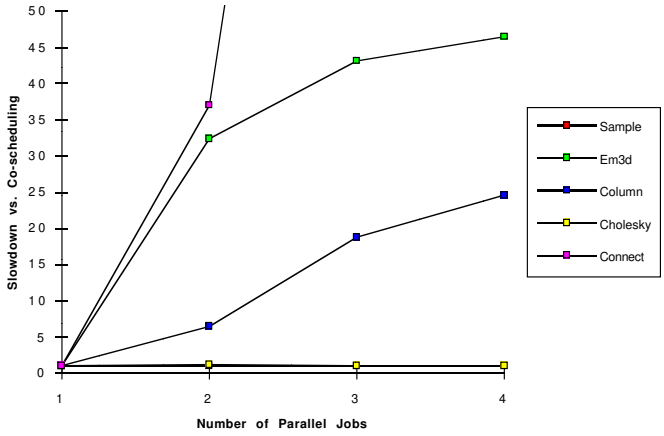
Machine (Procs)	***ODE***	***Transport***	***I/O ****	***Total* (s)**
Cray C-90 (16)	**7**	**4**	**25**	**38**
Intel Paragon (256)	**12**	**24**	**10**	**46**
NOW:				
RS/6000 (256) + PVM / Ether	**4**	**23,340**	**4,030**	**27,374**
+ ATM bandwidth (20 MB/s)	**4**	**192**	**2,015**	**2,211**
+ Parallel File System	**4**	**192**	**10**	**206**
+ low overhead (10 µs)	**4**	**11**	**10**	**25**

Order of Importance

*(1 disk / processor, parallel FS for C-90, Paragon)

Demmel and Smith, "Parallelizing a Global Atmospheric Chemical Tracer Model", SHPCC 1994.

Parallel Pgms like to Run in Parallel



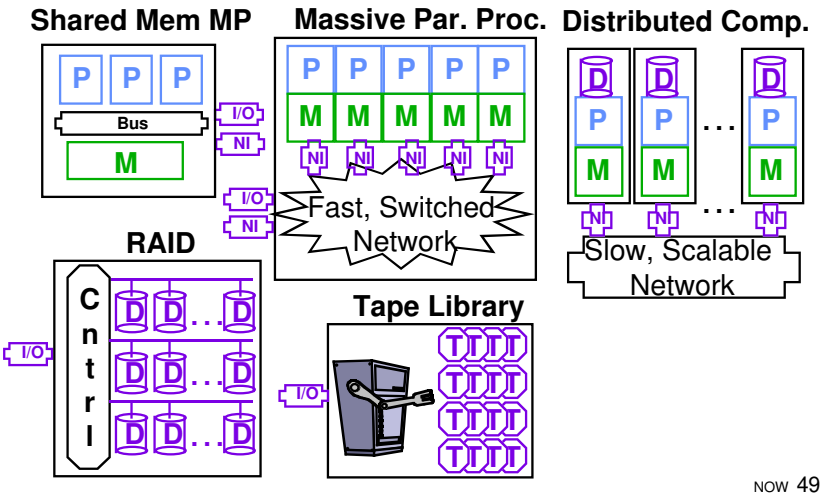
The Design of a Global O/S

- **How do you build it?**
  - **Out of off-the-shelf commercial components**
    - » **leverage huge investment / volume, reduce lag time, . . .**
    - » **technology transfer**
  - **Enabling technology**
    - » **Software fault isolation (SFI)**
    - » **Proxy**
- **What should it do?**
  - **Provide each active user with the interactive performance of one dedicated workstation**
  - **Provide a small number of users a large fraction of resources**
    - » **sequential or parallel**
  - **Be Available**

Preserving Interactive Performance

- **Pay attention to the state of the storage hierarchy**
  - **Dedicated workstation is aligned to "user's working set"**
    - » You notice intrusions that cause page faults
  - **Primary metric is the cost of restoration**
    - » **CPU is almost free**
    - » **Memory footprint is critical**
  - **Desktop: give back what you borrow**
  - **Basement: create affinity or lose**
- **Use Fast network and Parallel File System**
  - **Cost of restore**
    - » **64MB over Ethernet, single disk 60 seconds**
    - » **64MB over ATM, parallel file sys 4 seconds**

Desktop vs. Machine Room ?

- **BOTH !!!**
- **Desktop**
  - **Tremendous idle capacity, even at coarse grain**
  - **Very few unexpected reboots**
  - **Institutional infrastructure will support it**
  - **Want fast communication to my display**
- **Machine room**
  - **Simpler administration**
  - **Electrical vs optical**
  - **Always want some headroom in systems**
  - **Want baseline of resources for production**
- **Economics and workload will determine mix**

Taxonomy for Talk