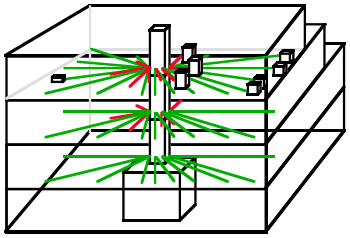
A Case for Networks of Workstations (NOW)

David Culler,
Tom Anderson,
Dave Patterson, et al.



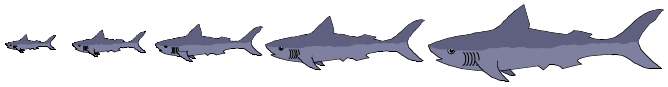
Computer Science Division
University of California, Berkeley

Mosaic: http://now.cs.berkeley.edu/ Related paper: IEEE Micro, Feb. 1995.

Outline

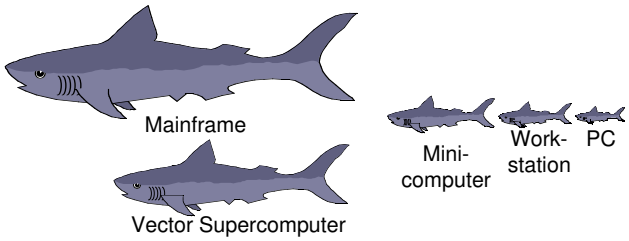
- **Why NOW now?**
  - **beyond the attack of the killer micros**
  - **emergence of the killer network**
- **Opportunities and Challenges**
  - **fast trusted-area communication**
  - **rethinking file systems and virtual memory**
  - **parallel computing**
  - **management of global resources**
- **Conclusion**

Biological Food Chain



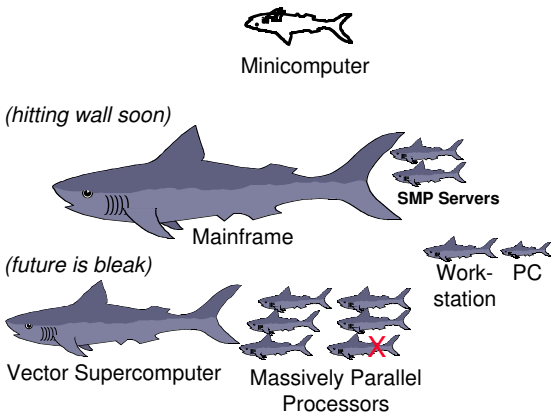
Case 3

1984 Computer Food Chain



Case 4

1994 Computer Food Chain



Case 5

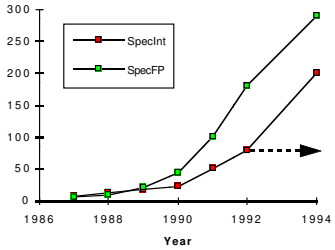
In the wake of the "killer micro"

+ **Can solve much bigger problems "on the desktop," but**
— **Large premium to solve any problem "bigger than the desktop,"**
  – **more compute, more memory, more disk**

Due to the cost of integration

Case 6

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
- MPPs typically have had a 2-3 year lag
- Servers usually 1-2 year lag

Case 7

Dis-Economy of scale

- Small (cheap) systems have large volume
- Learning curve: *"When volume doubles, costs reduce by 10%"**
  - at 1,000x volume => costs differ by 3x
- Extra engineering cost is amortized over fewer machines
  - same for sales, maintenance, . . .
- Differences in software costs >> HW factors

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

Case 8

Why NOW now?

- **Building block is big enough** (v. Intel 8086)
  - 1994 WS CPU ~ 1/3 supercomputer CPU
  - heavily under-utilized
- **Networks are faster**
  - Higher link bandwidth (v. 10 Mbit Ethernet)
  - Switch based networks coming (ATM, MPP backplanes)
  - New simple & fast paradigm for net. interfaces (Active Msgs)
- **I/O bottleneck**
  - Use network as I/O backplane (software RAID)
- **Demise of mainframes, supercomputers & MPPs**

- **Enables unique system design wins**
  - remote memory is much "closer" than local disk

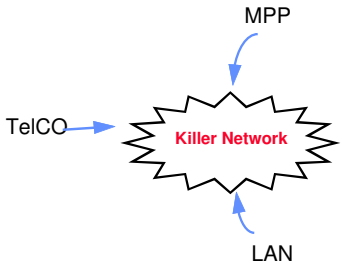
Case 9

Outline II

- **Why NOW now?**
  - beyond the attack of the killer micros
  - emergence of the killer network
- **Opportunities and Challenges**
  - fast trusted-area communication
  - file systems and virtual memory
  - parallel computing
  - management of global resources
- **Conclusion**

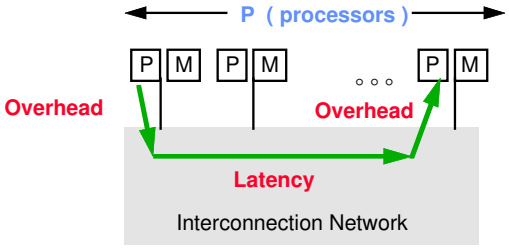
Case 10

Towards the Killer Network



Case 11

Communication Cost (beyond BW)

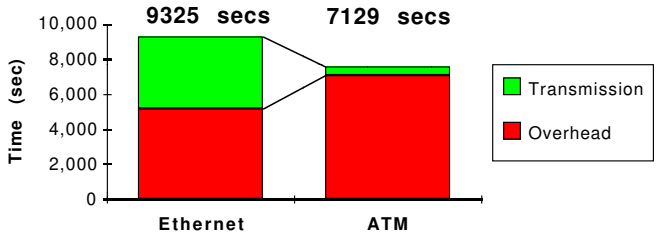


- **Network Latency incurred in sending message between nodes (1-way)**
- **Processor Overhead to send *or receive* a message**

Case 12

Importance of Overhead (+ Latency)

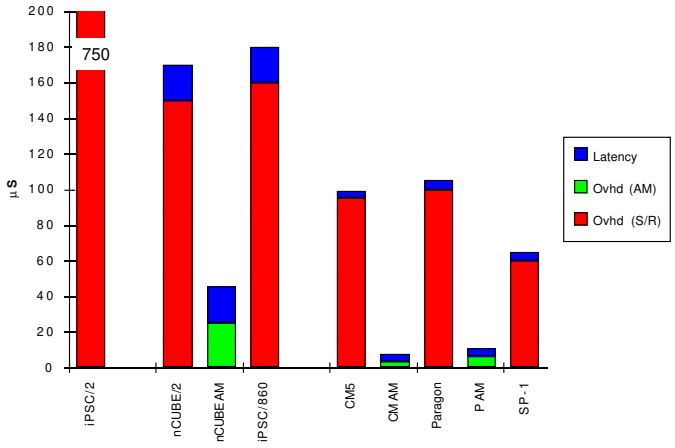
- **Ethernet / SS10: 9 Mb/s BW, 900 μsecs ovhd**
- **ATM Synoptics: 78 Mb/s BW, 1,250 μsecs ovhd.**
- **NFS trace over 1 week: 95% msgs < 200 bytes**



- **Link Bandwidth as misleading as MIPS**

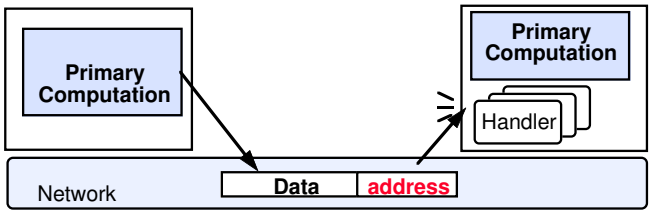
Case 13

MPPs: Low Latency HW is not Enough



Case 14

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.

Case 15

New Challenges in a NOW

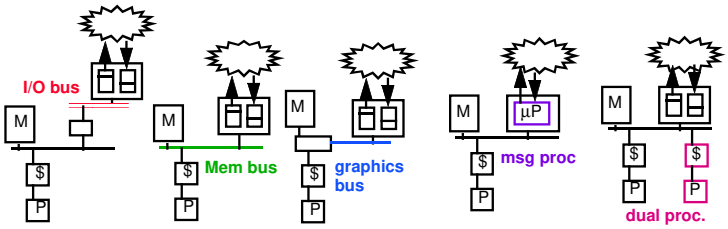
- **General Purpose Environment**
- **Serious Multiprogramming**
  - **Network Protection Domain**
  - **Naming**
  - **Resource Sharing**
  - **Demultiplexing input flows**
- **Full Function Operating System**
  - **page faults, daemons, . . .**
- **Packet Loss**
  - **infrequent** (if we do our job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

Traditional LAN: ~ 1,000 μs 1-way
NOW Goal: 10 μs 1-way

Case 16

Rough End of the Stone ?

- Rich HW Design Spectrum

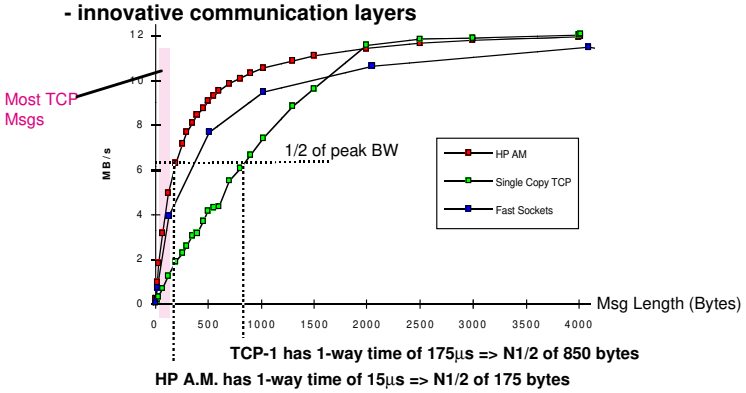


- Communication "Stack"
  - **simple protocols for "nice" networks**
  - **user level "virtual network"**

Case 17

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.

Case 18

Why it works

- Careful engineering of a simple abstraction
- Closer binding between network and user
  - external scheduler handles resource sharing
- Statefull NI Card + Request/Response permits very simple approach to
  - flow-control
  - time-out and retry

Much more to be done...

Case 19

More to be done...

- Network Protection Domain
  - NI inserts and validates “network PID”
  - Open portion of address space to comm. “partners”
- Naming of destination
  - Process routing
- Demultiplexing input flows
  - Multiple threads per node using “fast” communication
- Handling Lossage and Errors
  - Very low cost to prepare for very infrequent case

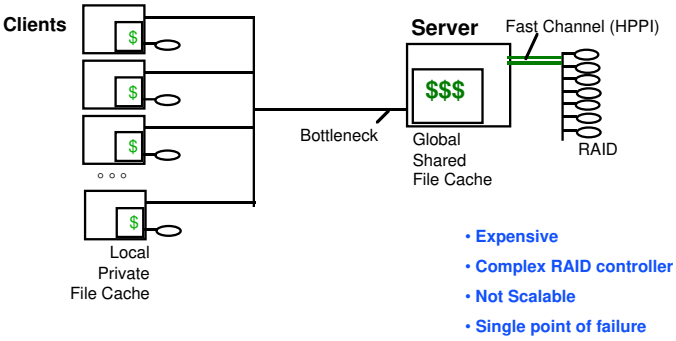
Case 20

Outline III

- **Why NOW now?**
  - **beyond the attack of the killer micros**
  - **emergence of the killer network**
- **Opportunities and Challenges**
  - **fast trusted-area communication**
  - **file systems and virtual memory**
  - **parallel computing**
  - **management of global resources**
- **Conclusion**

Case 21

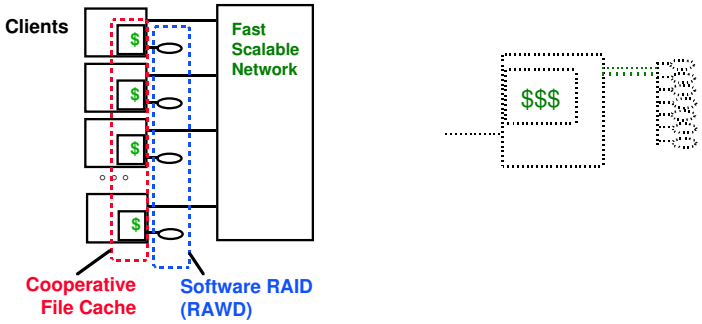
Dark Side of Client / Server



- **Server resources at a premium**
- **Client resources poorly utilized**

Case 22

NOW Client / Serverless File System



Client workstations cooperate to provide file service

Case 23

xFS (Anderson, et al)

- **Dynamic location of data, meta-data, and control**
  - **Easier migration, easier failure recovery**
- **Multiprocessor-style cache coherence**
  - **Maximize locality via write ownership**
  - **Direct client-to-client transfers**
  - **Latency critical**
- **Software RAID and Log Structured (LFS)**
  - **Parallel I/O, highly available despite client crashes**
  - **Network link bandwidth determines striping**
- **Cooperative caching**
  - **Aggregate client memory as giant cache for disk**
  - **client disks as giant cache for tape**

Case 24

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)

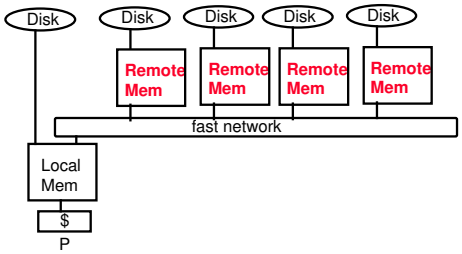
Dahlin, Wang, Anderson, and Patterson, "Cooperative Caching: Using Remote Client Memory to Improve File System Performance", OSDI 1994.

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**

Network Virtual Memory



4 ms + 4 KB / 2 MB/s = 6,000 µs

50µs + 4 KB / 20 MB/s = 250 µs

- **Potential 25x improvement in miss cost**
  - **Prototype user level pager: + 500 µs**
  - **Current VM systems: + 500 µs => demonstrate 5x**
- **Prototype network RAM distributed server**
  - **Invaluable insight toward communication abstractions**

Case 27

Outline IV

- **Why NOW now?**
  - **beyond the attack of the killer micros**
  - **emergence of the killer network**
- **Opportunities and Challenges**
  - **fast trusted-area communication**
  - **file systems and virtual memory**
  - **parallel computing**
  - **management of global resources**
- **Conclusion**

Case 28

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 FS	4	192	10	206
+ low overhead (10 µs)	4	11	10	25

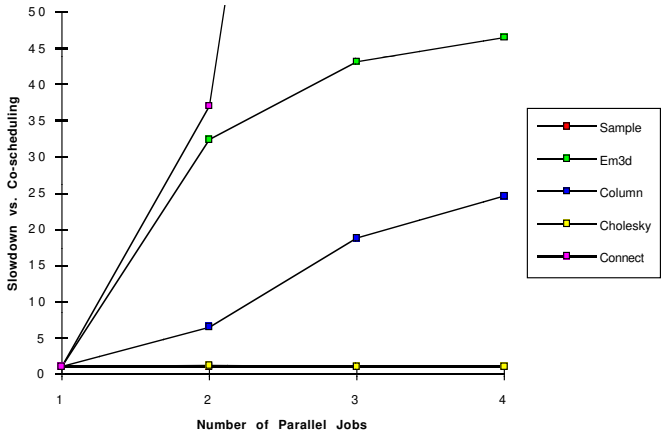
Order of Importance

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

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

Case 29

Parallel Pgms like to Run in Parallel



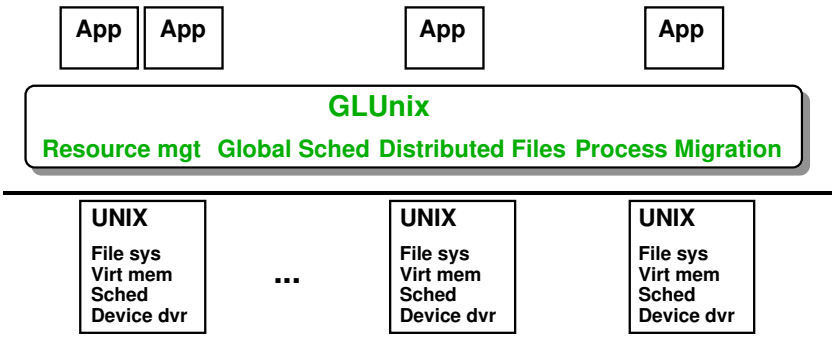
Case 30

Outline

- Why NOW now?
  - beyond the attack of the killer micros
  - emergence of the killer network
- Opportunities and Challenges
  - fast trusted-area communication
  - rethinking file systems and virtual memory
  - parallel computing
  - management of global resources
- Conclusion

Case 31

GLUnix: A Global Virtual O/S



GLUnix glues workstation UNIXs together to provide global services.

Case 32

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**

Case 33

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**

Case 34

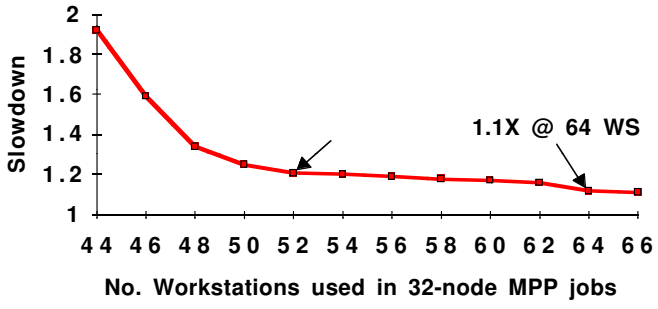
Demanding Applications

- **Sequential Applications**
  - network RAM
  - file system cooperative caching, disk striping
- **Parallel Applications**
  - co-scheduling
  - fast communication
  - parallel file systems
- **Availability**
  - engineering for fault isolation (not shared memory)
  - Hot Swap of HW and SW components
  - automated system management

Case 35

Sequential & Parallel on One System

- **Sequential has priority**
- **Ratio MPP nodes:desktops 3:5=>1.2x slowdown**



Arpaci, Dusseau, Vahdat, Lok, Anderson, Patterson. "The Interaction of Parallel and Sequential Workstations on a NOW," submitted.

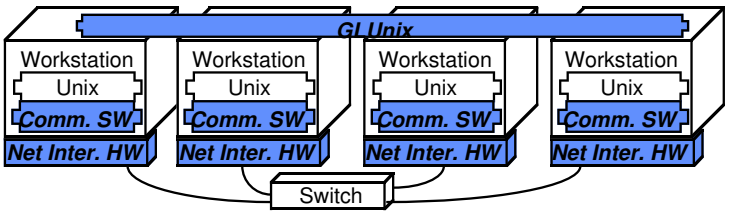
Case 36

Desktop vs. Machine Room ?

- My view: 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 adminstration
  - Electrical vs optical
  - Always want some headroom in systems
  - Want baseline of resources for production
- Economics and workload will determine mix

Case 37

NOW @ Berkeley



- Design & Implementation of higher-level system
  - Fast communication (low overhead SW / HW for Active Messages)
  - Global OS (Glunix)
  - Serverless Network File Systems (xFS)
- Goal: Better for Sequential and Parallel Appln's
  - recruit DRAM, disks, and CPUs
  - provide cheap, high availability
  - simplify system management

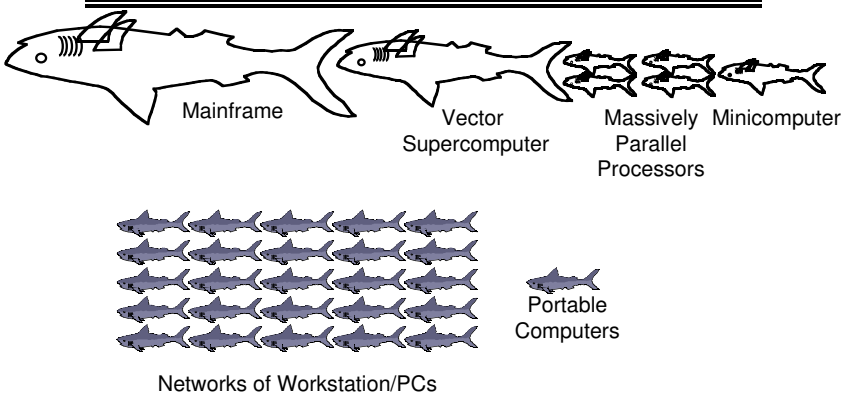
Case 38

Conclusion

- **Technologies aligned to exploit NOW now**
  - **32-bit μprocessors, switch based LANs, active messages, striped files, file caches, process migration**
- **1990s building block is desktop HW&SW**
  - **Need higher-level system research use building blocks: stand on shoulders, not toes**
- **Challenges for NOW**
  - **Leverage technology yet extend to make NOW ready**
  - **add low overhead user communication, global OS, serverless file system**
- **More than Parallel Computing**
  - **Make sequential users happy NOW**
    - » **recruit DRAM, disks, and CPUs**
    - » **provide cheap, high availability**

Case 39

2004 Computer Food Chain ???



Case 40