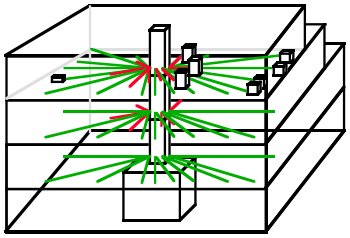
Research Agenda for NOW

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

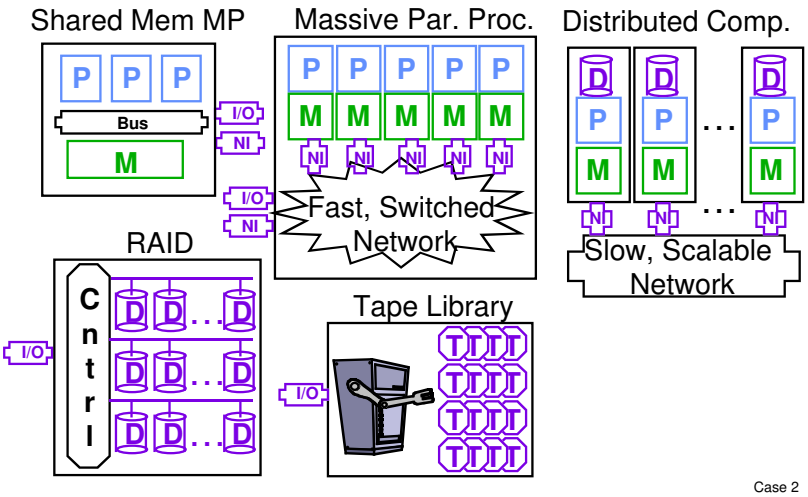
Computer Science Division
University of California, Berkeley



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

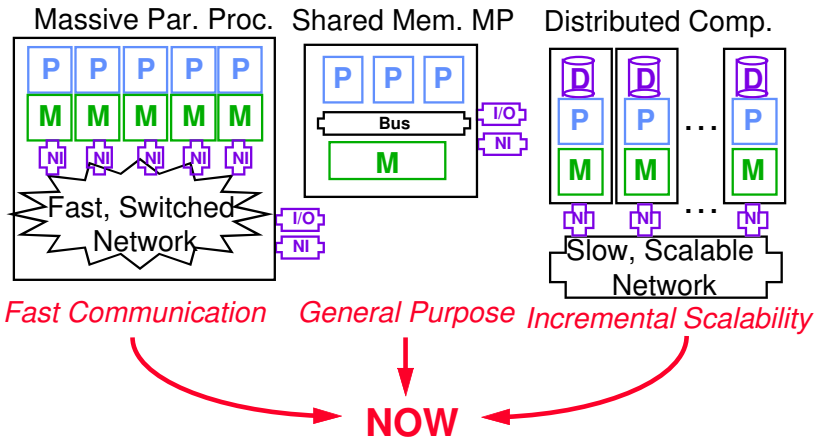
Case 1

Icons for Talk



Case 2

NOW Heritage, SMP/MPP/DC Future



Case 3

NOW: New Conceptual Framework

- **HW: Find Fast Changing Technology => reduce time delays => connect O.T.S. HW & SW**
- **SW: Offer Global Services on Local Devices**

Problem	**Big Iron Solution**	**NOW Solution**
WWW Search	**Disk based, SMP**	**RAM based, NOW**
Virtual Memory	**Local Disk**	**Remote RAM**
File Service	**SMP (Auspex)**	**Serverless FS**
Mass Storage	**RAID/Tape Library**	**Tertiary Disk**
High Speed Net	**Gigabit Network**	**Striped Network**
Number Crunch	**MPP, Vector SMP**	**NOW**

Case 4

NOW: Broad, Long Term Research Agenda

Narrow Agenda

- Fast Communication

Broad Agenda (new services/applications)

- Serverless file systems
- Network RAM
- WWW Search Engine

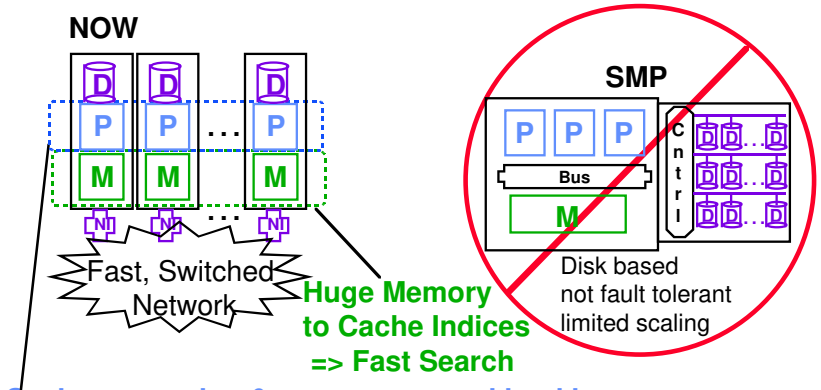
Long Term Agenda

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

...

Case 5

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



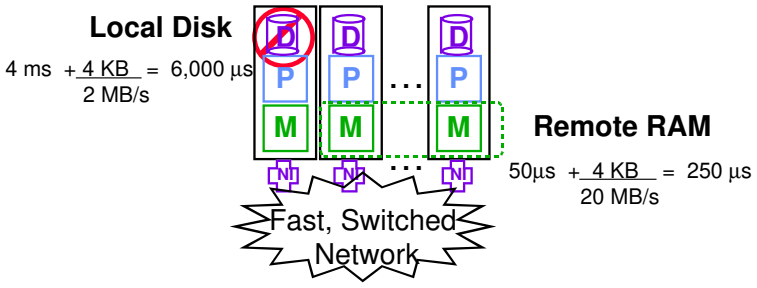
Scale processing & memory as workload increases

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

Graceful SW upgrade for 365x24 availability

Case 6

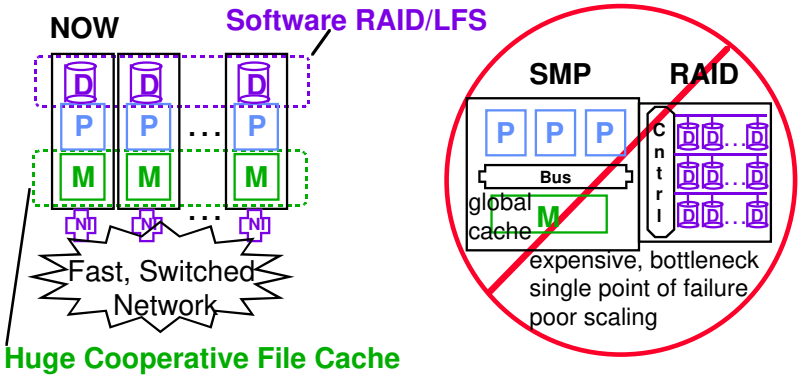
NOW Example: Fast Virtual Memory (Network RAM)



- **Current VM systems page to local disk => msec**
- **Instead, go over network to Remote RAM => µsec**

Case 7

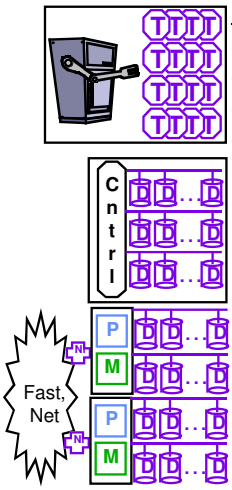
NOW Example: Truly Distributed F.S. (Serverless File Systems)



Huge Cooperative File Cache
- **private and global**
- **dynamic location (COMA + fault tolerance)**
- **latency critical**

Case 8

NOW Example: Mass Storage (Tertiary Disks)



Tape Library: Slow, Improving slowly
- Access time => secs, reading 10 TB => 1 week
- Tape $/MB improving 1.4X / year
- Disks improving $/MB by 2X / 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
- RAID systems cost ≈10X $/MB as a single PC disk (higher margin)

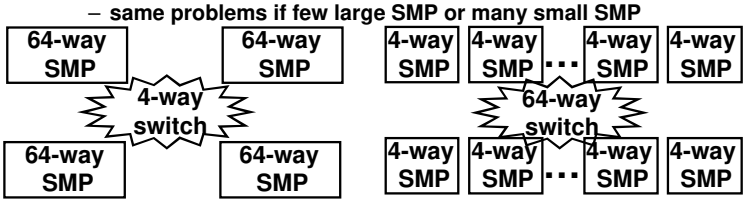
Tertiary Disk: Inexpensive Scaling
- O.T.S. racks, cables, computers, NIs, switches + SW RAID + maximum disks/computer
- Customer installs "just in time" disks, gets volume discount (≈25% @ 100)

$/MB,MTBF 1 disk same as 1000 disks

Case 9

Conclusion: Broad, Long Term Agenda

- **Future: Lots of people, Lots of computers**
  - some centralized, some on desk
- **Overwhelming desire for all computers 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 research challenges (HW &SW) will be in connecting boxes together**

Case 10

To Study Further

Mosaic: http://now.cs.berkeley.edu/

Anderson et al, “A Case for NOW”*IEEE Micro*, Feb. ‘95.

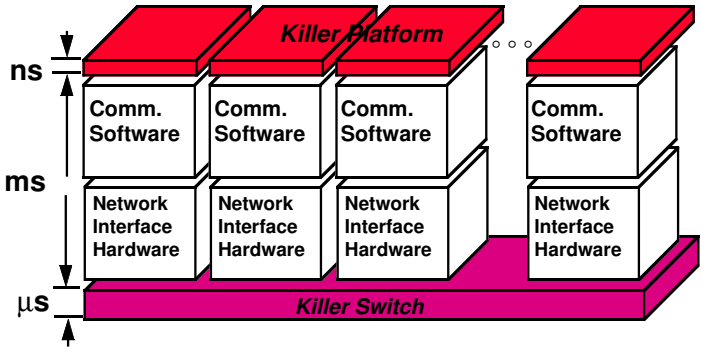
Arpaci et al “The Interaction of Parallel and Sequential Workstations on a NOW,” *SIGMETRICS*, ‘95.

Dahlin et al, “Cooperative Caching: Using Remote Client Memory to Improve File System Performance”,*OSDI* ‘94.

Martin “HPAM: An Active Message Layer for Networks of HP Workstations”, *Hot Interconnects* ‘94.

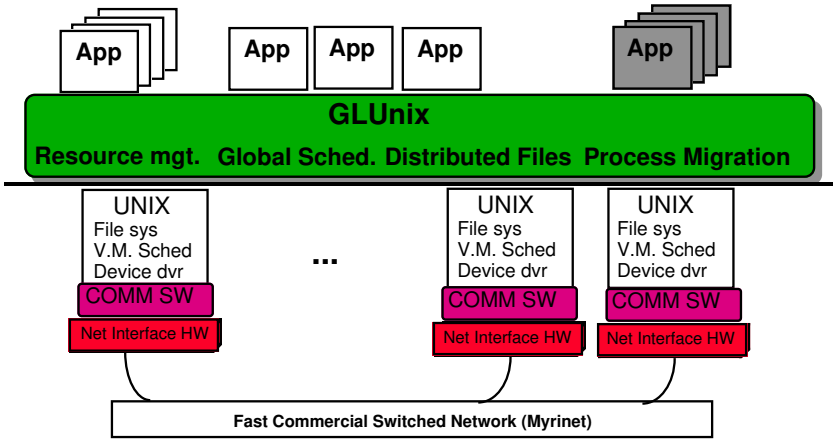
von Eicken et al. “Active Messages: A Mechanism for Integrated Communication and Computation”, *ISCA* ‘92.

Challenge: realizing fast communication



Goal: 10 µs end-to-end

Berkeley NOW System Architecture



Case 13

Project Components

- **Fast Trusted-Area Communications**
- **Serverless Parallel File Systems (I/O)**
- **Network Virtual Memory (Storage)**
- **Parallel Computing (Proc)**
- **Global Resource Management (O/S)**
- **Benchmarking and Evaluation**

Case 14