I/O Trends at NOAA/NCEP

George VandenBerghe
IMSG at NOAA/NCEP/EMC

November 6, 2008

www.imsg.com

Outline

- NCEP Overview & Driving CPU Trend
- Disk Trends
- Mass Store Trends

Overview

- NCEP is the United States’ premier NWP institution.
- First formed in 1954.
- Available compute of that time was a few kflops.
- Today it exceeds ten tflops.

NCEP COMPUTING

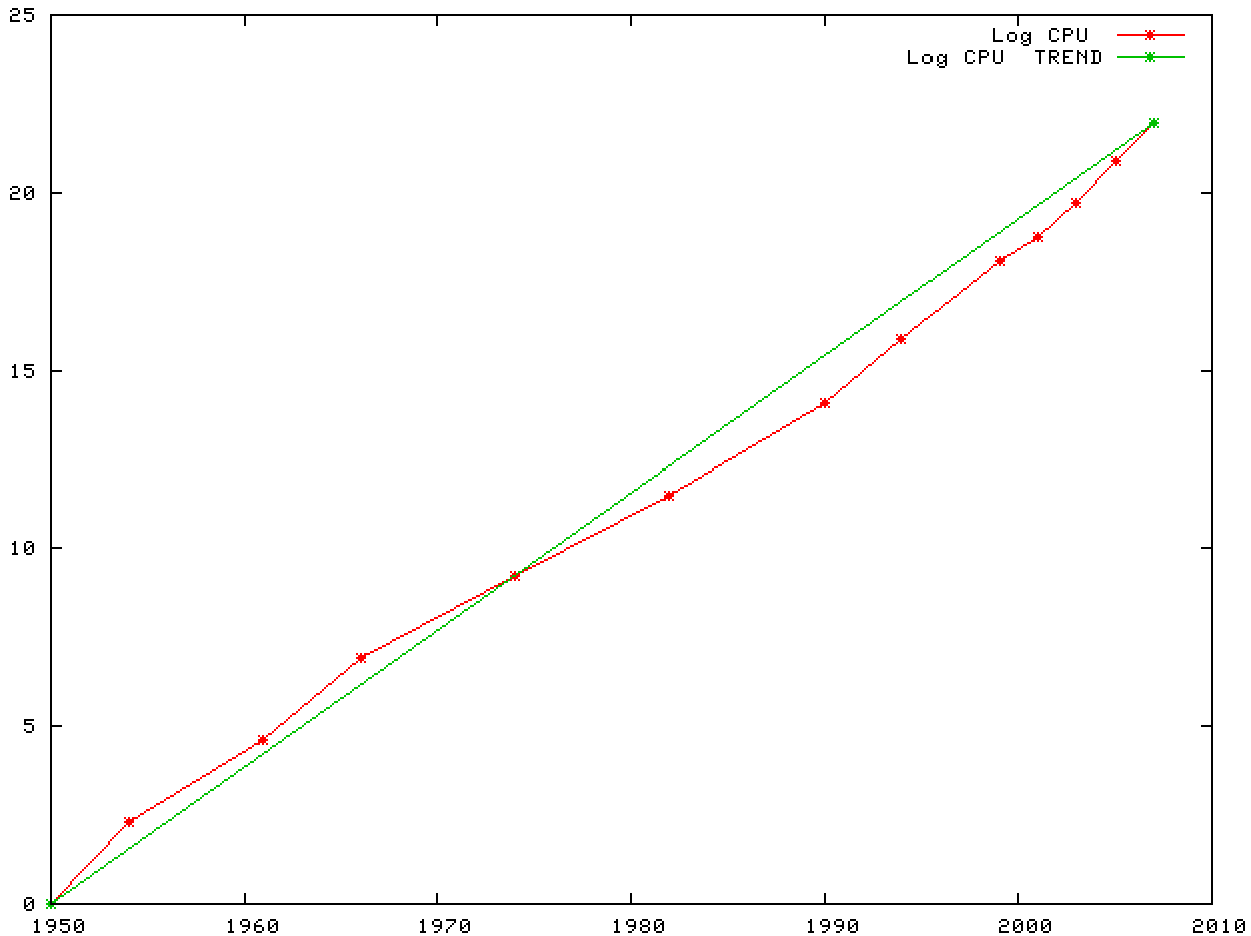
- There have been ~~28~~ ~~31~~ 32 doublings in compute capacity since 1950.
- The crossed out figures were from the Summer 2002 and 2006 Spscicomp presentations.
- Long term site is following “Moore’s Trend”
- Processor counts, flat for six years, have now doubled twice since 2003.

Past Platforms.

- ENIAC 1 kflop
- IBM 70x mid 50s 10 kflop
- IBM709x early 60s 100 kflop
- CDC6600 mid 60s-early 70s: 1 mflop
- IBM 360/195 x3 70s-early 80s 10 Mflop
- CDC CYBER205 x2 80s 100 Mflop
- CRAY Y-MP8 early 90 1Gflop (1.3 aggregate)
- CRAY C90 mid to late 90s 6 Gflop (8 aggregate)
- IDM SP 1999-2000 30 gflop (70 aggregate)
- IBM SP 2000-2002 60 gflop (140 aggregate)(x2)
- P690 (P4) 2003-2004 160 gflop (370 aggregate)(x2)
- P655 (P4+) 2005-2006 500 gflop (1150 aggregate)(x2)
- P575 (P5) 2007-2008 1500 gflop (3400 aggregate)(x2)
- Another tripling from P6 coming soon.
- ***Available cycles have doubled every two years since 1950 with more rapid doubling in recent two decades*** ***(correlation with my career is coincidental).***

Past Platforms.

- ENIAC 1 kflop
- IBM 70x mid 50s 10 kflop
- IBM709x early 60s 100 kflop
- CDC6600 mid 60s-early 70s: 1 mflop
- IBM 360/195 x3 70s-early 80s 10 Mflop
- CDC CYBER205 x2 80s 100 Mflop
- CRAY Y-MP8 early 90 1Gflop (1.3 aggregate)
- CRAY C90 mid to late 90s 6 Gflop (8 aggregate)
- IDM SP 1999-2000 30 gflop (70 aggregate)
- IBM SP 2000-2002 60 gflop (140 aggregate)(x2)
- P690 (P4) 2003-2004 160 gflop (370 aggregate)(x2)
- P655 (P4+) 2005-2006 500 gflop (1150 aggregate)(x2)
- P575 (P5) 2007-2008 1500 gflop (3400 aggregate)(x2)



A FEW SITE TRENDS (2002)

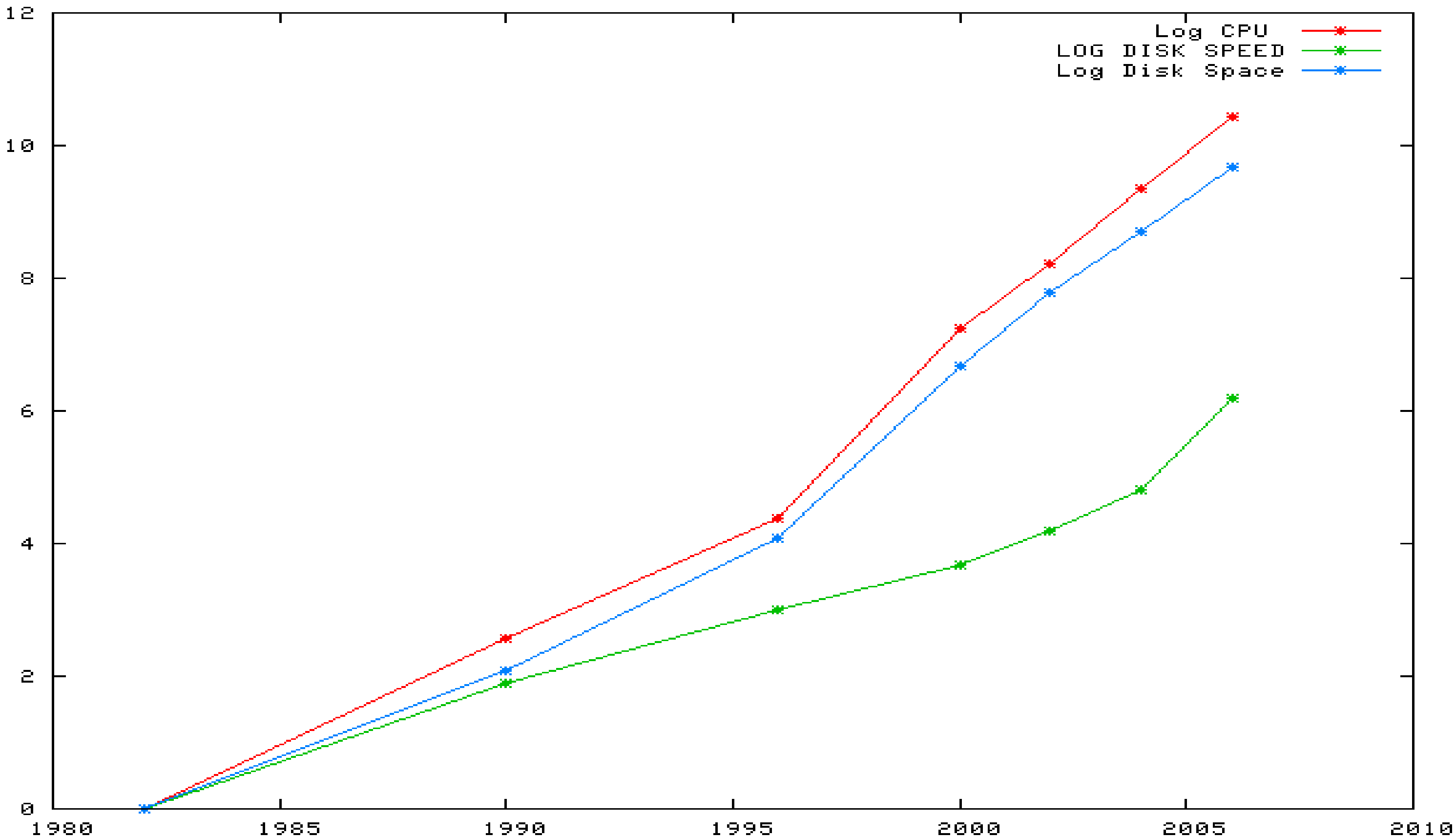
- Compute requirements grow exponentially.
- E-fold time ~2 years.
- Compute budget near constant. (e-fold time 20+ years)
- Watch out for per flop costs that are constant or decreasing with e-fold time >2 years)
- (floor space/tflop, disk/tflop, tapes/tflop)
- We’re okay for disk/tape capacities and floor space but not performance.

Disk Farm Service Metrics.

- 1982, Cyber 205 15 mbytes/second 10GB
- 1990 Cray YMP 100 mbytes/second 80GB
- 1996 Cray C90 300 mbytes/second 600GB
- 2000 IBM/SP 600 mbytes/second 16TB
- 2002 P690 1000 mbytes/second 24TB
- 2004 P655 2000 mbytes/second 60TB
- 2006 P575 8000 mbytes/second 160TB

Disk Speed & Space V/S CPU.

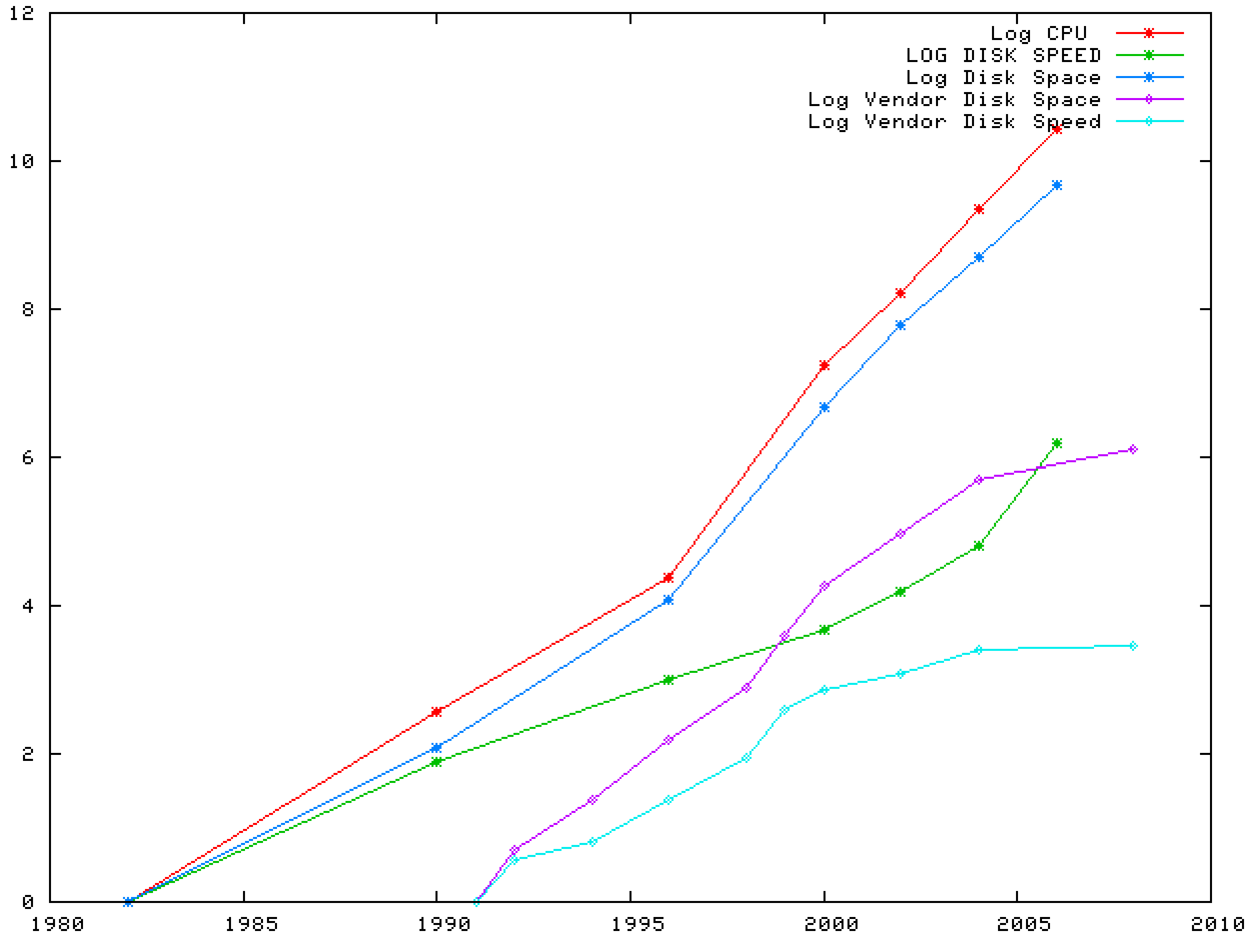
- 1982, Cyber 205 15 mbytes/second 10GB
- 1990 Cray YMP 100 mbytes/second 80GB
- 1996 Cray C90 300 mbytes/second 600GB
- 2000 IBM/SP 600 mbytes/second 16TB
- 2002 P690 1000 mbytes/second 24TB
- 2004 P655 2000 mbytes/second 60TB
- 2006 P575 8000 mbytes/second 160TB



Disk Speed & Space V/S CPU.

- 1982, Cyber 205 15 mbytes/second 10GB
- 1990 Cray YMP 100 mbytes/second 80GB
- 1996 Cray C90 300 mbytes/second 600GB
- 2000 IBM/SP 600 mbytes/second 16TB
- 2002 P690 1000 mbytes/second 24TB
- 2004 P655 2000 mbytes/second 60TB
- 2006 P575 8000 mbytes/second 160TB

Two new curves are vendor disk speeds and capacities from 1991 to present. Space and speed match the other curves.
(Source Henry Newman Instrumental Inc.)



I/O Issues

- Disk space/flop has scaled linearly.
- Disk performance per flop *has not!*
- This long going IT trend was masked by low I/O requirements of traditional NWP.
- Large deterministic forecast problem sizes grew more slowly than cpu capacity because of CFL constraints.
- Ensemble requirements scale linearly with cpu capacity and transactions/second also scales linearly.
- I/O desires are now a linear function of compute capacity. (I ~ k*C where C is compute)

I/O

- Disk metric trends are MUCH flatter than compute. (Both transactions/sec and bandwidth)
- These are partially covered by device parallelism (striping or multiple independent disks or a combination of both).
- Single disk speed ratio 1982:2006 is about 3:60 (mbytes/sec) (factor of 20)
- Single stream ratio is 3:300 (factor of 100)
- Multiple aggregate stream is 15:8000 (factor of 530)
- Compute ratio is 100:3,400,000 (Factor of 34,000)
- Ratios are from 1982 Cyber 205 V.S. 2006 P575.

Maximum Rate with JBOD

- We have 1200 disks.
- A recent timing showed 60mb/sec each.
- Absolute max throughput is 72GB/sec if enough parallel machines, cables, adapters, etc. are involved.
- Max throughput on C205 disks with enough controllers was 48MB/sec.
- Hardware speedup is 1500x.

Bare Metal I/O

- Single disk speeds have increased by 20x. (60/3)
- Single disk TPS has increased by 10x (optimistic estimate)
- Single disk capacity has increased by 200x. (133/0.6)
- If capacity tracks cpu capacity then TPS and bandwidth lags by 10/200 and 20/200 assuming OPTIMAL hardware layout (no filesystem overhead or network overhead or raid overhead)

Ratios

- P575/C205 compute 34000x
- P575/C205 agg. I/O 8000/15 or 533x
- P575/C205 *theoretical* hardware agg. 72000/48 (1500x)
- Capacity/Speed ratio argument 200x/20x.(10x less)
- P575 aggregate/flop 63x less
- P575 hardware max aggregate/flop 22x less
- .We might get 2-3x improvement with increasingly expensive I/O system and filesystem accommodations.
- Numbers in red should be the same. However our capacity/flop has slipped slightly (16000/34000) or factor of 2.1 less. Difference between 2.1 and 2.2 is due to rounding.

Ratios

- Capacity/Speed ratio argument 200x/20x.(10x less)
- P575 aggregate/flop 63x less
- P575 hardware max aggregate/flop 22x less

- So we've lost a factor of 60 in aggregate disk farm bandwidth.
- We could get a factor of 2 back by doubling disk infrastructures (probably just doubling disks and filesystem count from 1 huge one to 2). In short, doubling capacity.
- We could get another factor of 3 back with more disk servers, networks, cables, larger numbers of independent filesystems, and other questionable tradeoffs.
- (We could get another factor of 3 by somehow eliminating all filesystem, controller , connectivity and server overheads but these aren't even questionable, they're impractical)

- The future trend will be to buy excess capacity to get needed bandwidth *but we're not there yet.*

I/O

- The least ominous number for us is the bandwidth/capacity number.
- This is 10x less than in 1982.

- No matter how clever we are in microcode, firmware or configuration we can't get past this number! BUT
- They can be addressed in high level software design (do less I/O!, implement caches, replace files with pipes, etc.)
- *They may require paradigm shifts in thinking! BUT*
- *We can do 18x of increasingly difficult improvements to get to this number and 6x are reasonable.*
- *We do have time (several years) to do this.*
- *We need to anticipate this and adapt rather than encounter it and react.*
- **Next slides are examples of two easy to anticipate issues that were missed**

A Mild HPC Paradigm Failure

- Memory swapping/Job Roll.
- Common in 1970s and persisted in HPC longer than on smaller computers (Cray esp.) into late1990s
- Entire process memory moves to disk and is replaced with another to time slice memory)
- This worked in 1980 (1 mbyte process 3mb/sec disk, 0.6 second swap, repeat every minute: small impact)
- Similar thinking in 1992 (300 mbyte process 20 mbyte/second disk, 30 second swap, repeat every minute: Huge Impact!).
- In early 90s, most sites one way or another, configured to avoid swapping or swap to Solid State Devices (high latency memory).
- *This problem was not anticipated but was instead independently analyzed and solved at most sites.*

Memory Management Round 2

- Swapping big iron was gradually replaced with DSM Unix nodes in late 90s.
- These mostly PAGED rather than swapped.
- Paging issues in early 00s very similar to Swapping issues in early 90s.
- 1992, 16mb memory 1mb/sec page device, page 1/10 memory in 1.6 seconds. (one way)
- 2006 32GB memory 40mb/sec device 1/10 of memory paged in 80 seconds!
- In early to mid 00s we're avoiding, blocking, or cancelling paging processes. We can't tolerate these times! (in addition vital kernel and daemon services are blocked during heavy paging)
- Again this is being done in the field rather than from vendor recommendations!

And Speaking of Memory

- 1982 C205 32mbytes
- 2008 P575 node 32 gbytes
- 1000x increase in memory
- 100x increase in cpu/node
- Flush memory to single disk 1982 10 sec 2007 500 sec.
- Flush ½ memories (16G*160 nodes) to filesystem 1982 2 sec   2007 320 seconds

Disk Farm Service Metrics.

- 1982, Cyber 205 15 mbytes/second
- 1990 (10x) Cray YMP 100 mbytes/second (6.6x)
- 1994 (50x) Cray C90 300 mbytes/second(20x)

- 2000 (1400x) IBM/SP 600 mbytes/second (40x)
- 2002 (3700x) P690 1000 mbytes/second (66x)
- 2005 (11500x) P655 2000 mbytes/second (133x)
- 2007 (34000x) P575 8000 mbytes/second (532x)

- *Disk aggregate bandwidth/flop decreased by factor of 63 from 1982 to 2007 systems*.
- Disk space increased by 16000x over this period (space/flop decreased by a factor of only 2.1)

Mitigation Methods
Write Buffering.

- Single users can hide I/O with system or user specified buffers. (A special case of this is single I/O tasks which use task memory as a large buffer, gather from all other tasks and write offline.)
- This method is effective for single streams.
- INEFFECTVE against aggregate I/O deficiency.
- When aggregate write rates approach system maximum, buffers fill and I/O backs up.
- “Hockey Stick” time profile is typical, small changes in aggregate I/O or system service capability. produce changes in overhead from near zero to “large”.
- These methods increase user throughput and reduce runtimes but their breakdown is very rapid and analyst time to develop a mitigation strategy is much reduced.

- Very popular at NCEP but we were ready for it.

Other Mitigations

- Typical NCEP pattern is
- Forecast Write → post process → write → product generator → write.
- This generates many writes followed by reads followed by more writes.
- Alternative is to make post processor and perhaps product generator a part of the forecast model. NCEP is at an intermediate stage there.
- Disk I/O is replaced by interconnect transfers which are much faster and cheaper.

Mass Store (Tapes)

- 1982 Cyber 205, 250kb/sec to 1.6TB of round tapes
- 1990 Cray Y-MP 600kb/sec to 2TB of cartridge tapes
- 1993 Cray Y-MP 10mb/sec to 2TB of online tapes
- 1996 Crays 30mb/sec to 6TB of online tapes
- 1997 Crays 50mb/sec to 300TB of online Helical Scan tapes
- 2000 IBM/SP 40mb/sec to 400TB of offline Dual Copy TSM tapes
- 2003 P690 150mb/sec to 3PB of offline Dual Copy HPSS tapes
- 2007 P575 200mb/sec to 5PB of offline mostly single copy tapes
- 2008 400 mb/sec to 15 PB of offline tapes.
- 2008/1982 compute increase 34000x, Tape space 9000x, Tape speed 1600x, Compute/Space increase 3.7x. Compute/Speed increase 21X.

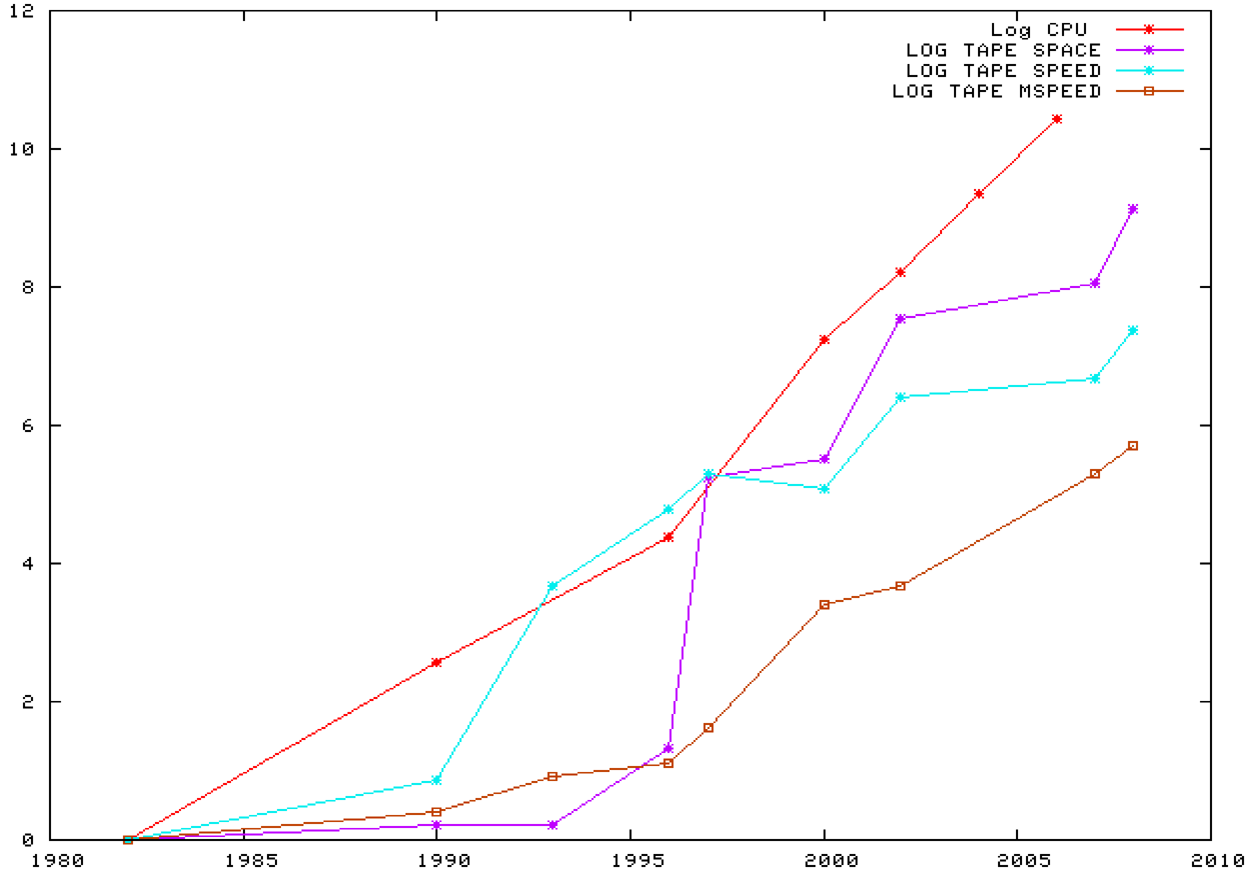
Mass Store (Tapes) (Hardware Max)

- 1982 Cyber 205, 10mb/sec to 1.6TB of round tapes
- 1990 Cray Y-MP 15mb/sec to 2TB of cartridge tapes
- 1993 Cray Y-MP 25mb/sec to 4TB of online tapes+cartridge
- 1996 Crays 30mb/sec to 6TB of online tapes
- 1997 Crays 50mb/sec to 300TB of online Helical Scan tapes
- 2000 IBM/SP 300mb/sec to 400TB of offline Dual Copy TSM tapes
- 2003 P690 400mb/sec to 3PB of offline Dual Copy HPSS tapes
- 2007 P575 2000? mb/sec to 5PB of offline mostly single copy tapes
- 2008 3000? mb/sec to 15 PB of offline tapes.
- 2008/1982 compute increase 34000x, Tape space 9000x, Tape speed 300x, Compute/Space increase 3.7x. Compute/Speed increase 113X.

- Difference between this and previous slide is this one assumes optimal hardware connections (direct connect 1 mover per drive). Red numbers are/were changeable with inexpensive hardware increments or configuration.
- A configuration where support increment cost exceeds drive count increment cost is not “optimal”

Mass Store (Tapes) (Hardware Max)

- 1982 Cyber 205, 10mb/sec to 1.6TB of round tapes
- 1990 Cray Y-MP 15mb/sec to 2TB of cartridge tapes
- 1993 Cray Y-MP 25mb/sec to 4TB of online tapes+cartridge
- 1996 Crays 30mb/sec to 6TB of online tapes
- 1997 Crays 50mb/sec to 300TB of online Helical Scan tapes
- 2000 IBM/SP 300mb/sec to 400TB of offline Dual Copy TSM tapes
- 2003 P690 400mb/sec to 3PB of offline Dual Copy HPSS tapes
- 2007 P575 2000? mb/sec to 5PB of offline mostly single copy tapes
- 2008 3000? mb/sec to 15 PB of offline tapes.
- 2008/1982 compute increase 34000x, Tape space 9000x, Tape speed 300x, Compute/Space increase 3.7x. Compute/Speed increase 113X.
- Difference between this and previous slide is this one assumes optimal hardware connections (direct connect 1 mover per drive). Red numbers are/were changeable with inexpensive hardware increments or configuration.

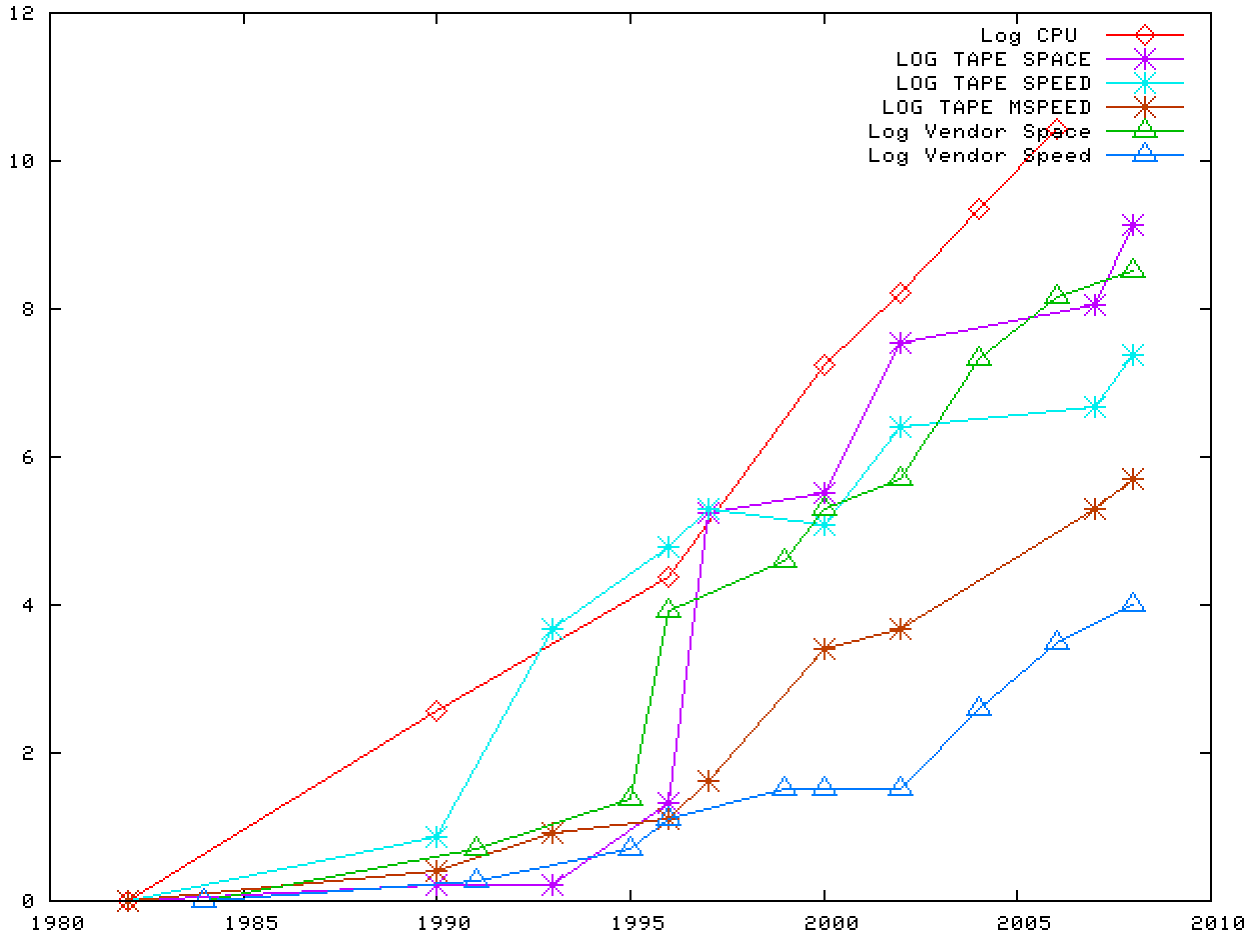


Mass Store (Tapes) (Hardware Max)

- 1982 Cyber 205, 10mb/sec to 1.6TB of round tapes
- 1990 Cray Y-MP 15mb/sec to 2TB of cartridge tapes
- 1993 Cray Y-MP 25mb/sec to 4TB of online tapes+cartridge
- 1996 Crays 30mb/sec to 6TB of online tapes
- 1997 Crays 50mb/sec to 300TB of online Helical Scan tapes
- 2000 IBM/SP 300mb/sec to 400TB of offline Dual Copy TSM tapes
- 2003 P690 400mb/sec to 3PB of offline Dual Copy HPSS tapes
- 2007 P575 2000? mb/sec to 5PB of offline mostly single copy tapes
- 2008 3000? mb/sec to 15 PB of offline tapes.
- 2008/1982 compute increase 34000x, Tape space 9000x, Tape speed 300x, Compute/Space increase 3.7x. Compute/Speed increase 113X.

- Difference between this and previous slide is this one assumes optimal hardware connections (direct connect 1 mover per drive). Red numbers are/were changeable with inexpensive hardware increments or configuration.

Dark blue and Green Curves are one vendor's published
device metrics since 1984 (source Henry Newman Instrumental Inc.)

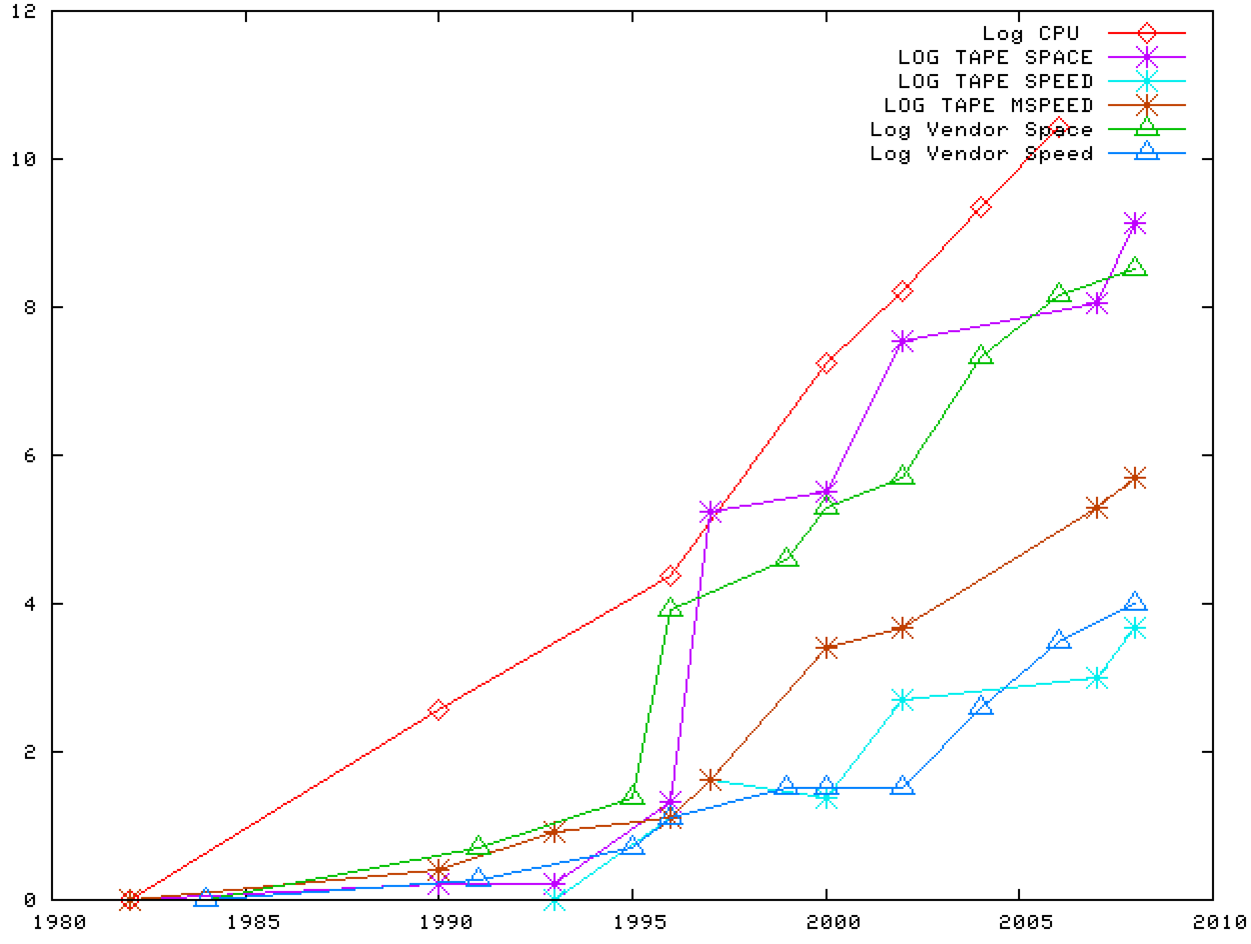


Mass Store (Tapes) (Hardware Max)

- 1982 Cyber 205, 10mb/sec to 1.6TB of round tapes
- 1990 Cray Y-MP 15mb/sec to 2TB of cartridge tapes
- 1993 Cray Y-MP 25mb/sec to 4TB of online tapes+cartridge
- 1996 Crays 30mb/sec to 6TB of online tapes
- 1997 Crays 50mb/sec to 300TB of online Helical Scan tapes
- 2000 IBM/SP 300mb/sec to 400TB of offline Dual Copy TSM tapes
- 2003 P690 400mb/sec to 3PB of offline Dual Copy HPSS tapes
- 2007 P575 2000? mb/sec to 5PB of offline mostly single copy tapes
- 2008 3000? mb/sec to 15 PB of offline tapes.
- 2008/1982 compute increase 34000x, Tape space 9000x, Tape speed 300x, Compute/Space increase 3.7x. Compute/Speed increase 113X.
- Difference between this and previous slide is this one assumes optimal hardware connections (direct connect 1 mover per drive). Red numbers are/were changeable with inexpensive hardware increments or configuration.

Dark blue and Green Curves are vendor's published device metrics since 1984 (sourc Henry Newman Instrumental Inc

As in previous slide But with site realized tape speed normalized to 1993 value. Previous speed was high because of a grossly suboptimal network interface

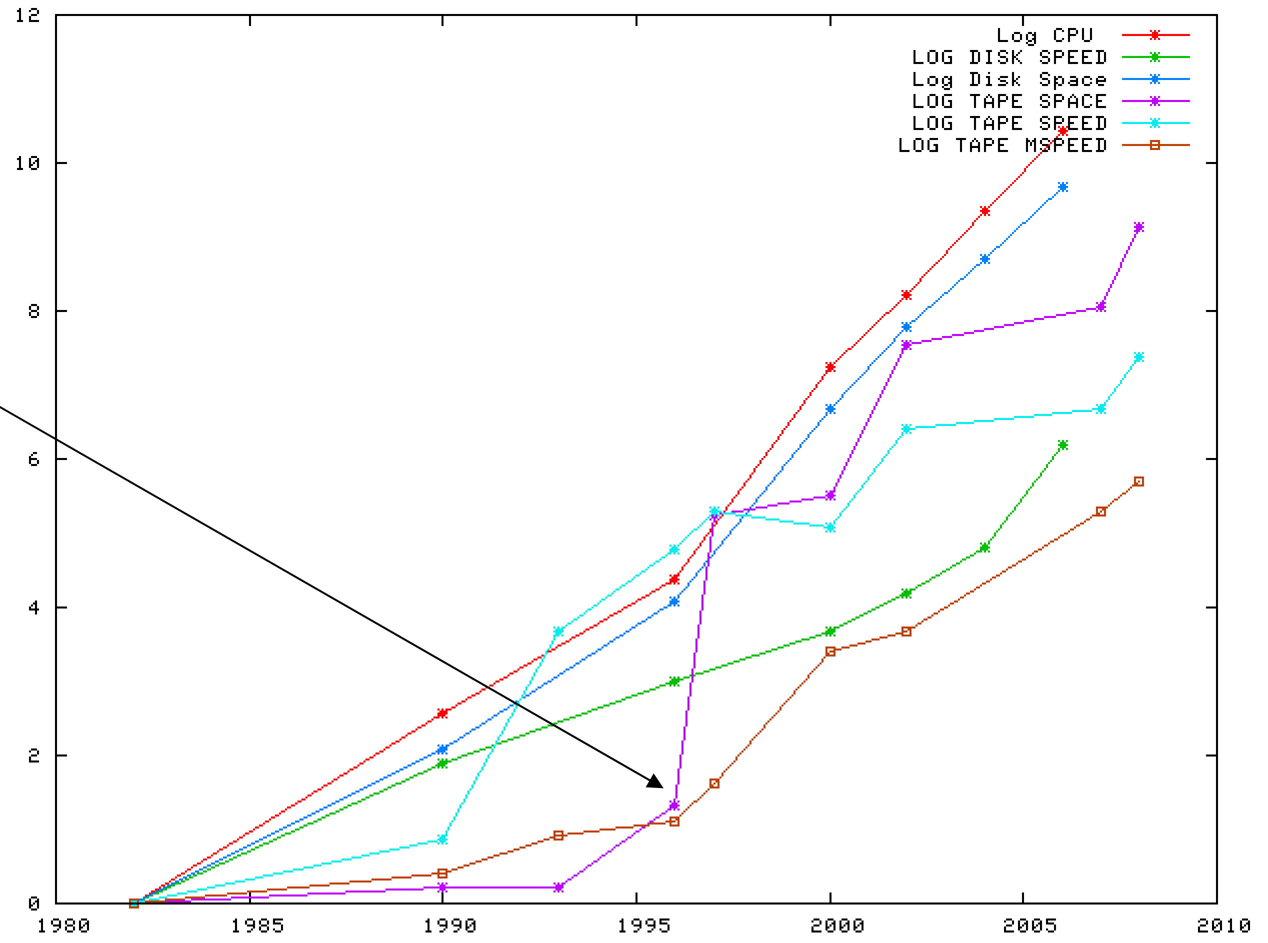


Bandwidth Conclusions

- There is more room to architect faster tape infrastructure (more network, and movers)
- In 2008 this could have gotten us another 6-7x speedup.
- We still need to watch this trend.

Tape And Disk V/S CPU

Huge density increase with STK“redwood”™ + Eagle™. Followup 9940B was not helical scan but was very high density)



Sample User Getaround Procedure

- With HPSS writes go to disk and then offline to tape.
- Write rate of X requires disk rates of 2X for the write+read.
- Usual recall policy is stage to disk then stream to network.
- *Streams from tape (configurable by COS in HPSS) are much faster and also save stage time.*
- Our tape cloud is faster than the movers' supporting disk pool aggregate rate.   Disk bandwidth is a scarce resource in the mover cloud.
- *So go against paradigm of staging from "slow" tape to "fast" disk and stream much faster directly from tape*

Write Expectations

- 7.6PB Year (FY 2009)
- 11.6PB (FY2010)
- Compare the FY2009 write rate of 240mbytes/sec with the installation I/O capability of 400mbytes/sec. 24x7 write stream by itself takes a large fraction of it!!
- Fortunately, expensive drives aren't the problem, it's disk and mover infrastructure in front of them which are less expensive.

Tape Transactions

- This situation is far more serious.
- 1990 10 drives 120 mounts/hour
- 2008 24 drives 720 mounts/hour
- 34000x increase in compute, 6x increase in mounts.
- Need to get a lot for every mount (very large files AND problem locality needed)
- Disk and movers and network won't help the transaction problem. If we can't get big files and locality, we need more (expensive) drives.

Problem We Don't Have

- Media density has tracked cpu capability with time.
- This means both floor space and tape room (silo) volumes to support disks and tapes are NOT increasing.

Conclusions

- Both Disk and Tape bandwidth are slipping by large factors relative to compute capacity. (63x and 21x(113x))
- Significant (not necessarily easy) opportunities still exist in optimizing filesystems, and tape support infrastructure before buying excess capacity or more drives to get bandwidth. (7.5x for tapes,9x for disks).
- The crunch argued in the abstract is coming but not as soon as earlier thought.
- Disk Space/Flop is slipping slightly (2x)
- Transactions to tapes need to be examined *very* closely.
- Assumption that we have to buy extra disk capacity to get enough bandwidth is not valid (yet!)

Questions??