Big Data ...

and the Next Wave of InfraStress

John R. Mashey

Chief Scientist, SGI

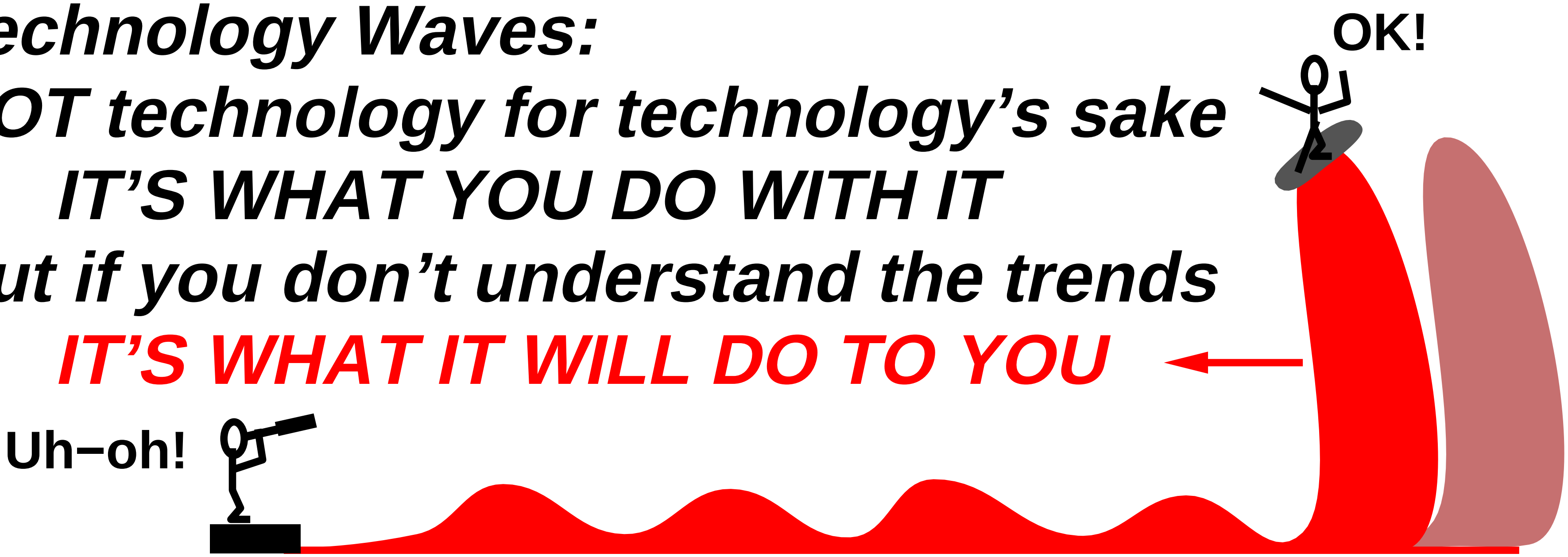
Technology Waves:

NOT technology for technology's sake

IT'S WHAT YOU DO WITH IT

But if you don't understand the trends

IT'S WHAT IT WILL DO TO YOU



Big Data

And The Next Wave of InfraStress

1. *Big data:* storage growing bigger faster
   - DRAM: 1.6X/year (4X/3 years) continues
   - Disk density:
   - 1.3X/year CAGR: historical trendline
   - 1.6X/year since ~1990
   - 2.0X/year leap ~1998/1999
2. *Net* continues raising user expectations
   - More data (image, graphics, models)
   - (Some) more difficult data (audio, video)
   - Pressure on net, especially last mile

=> Explosion of WIDELY−accessible data

Create, understand, store, move ... or else ...

Drown in Wave of Infrastructure Stress

General references: John Hennessy, David A Patterson, Computer Architecture: A Quantitiative Approach, Second Edition, Morgan Kaufmann, San Francisco, 1996. ISBN 1−55860−329−8.
Also, Computer Organization and Design, Morgan Kaufmann, San Francisco, 1994, ISBN 1−55860−281−X.
Also, thanks to Glenn Stettler of SGI, "Disk Drive Futures", 1/20/99.

InfraStress
= Infrastructure Stress

in·fra·stress. *n.*

1. Bad effects of faster change in computer *subsystems & usage*:

CPUs, memory, disks, demand ...

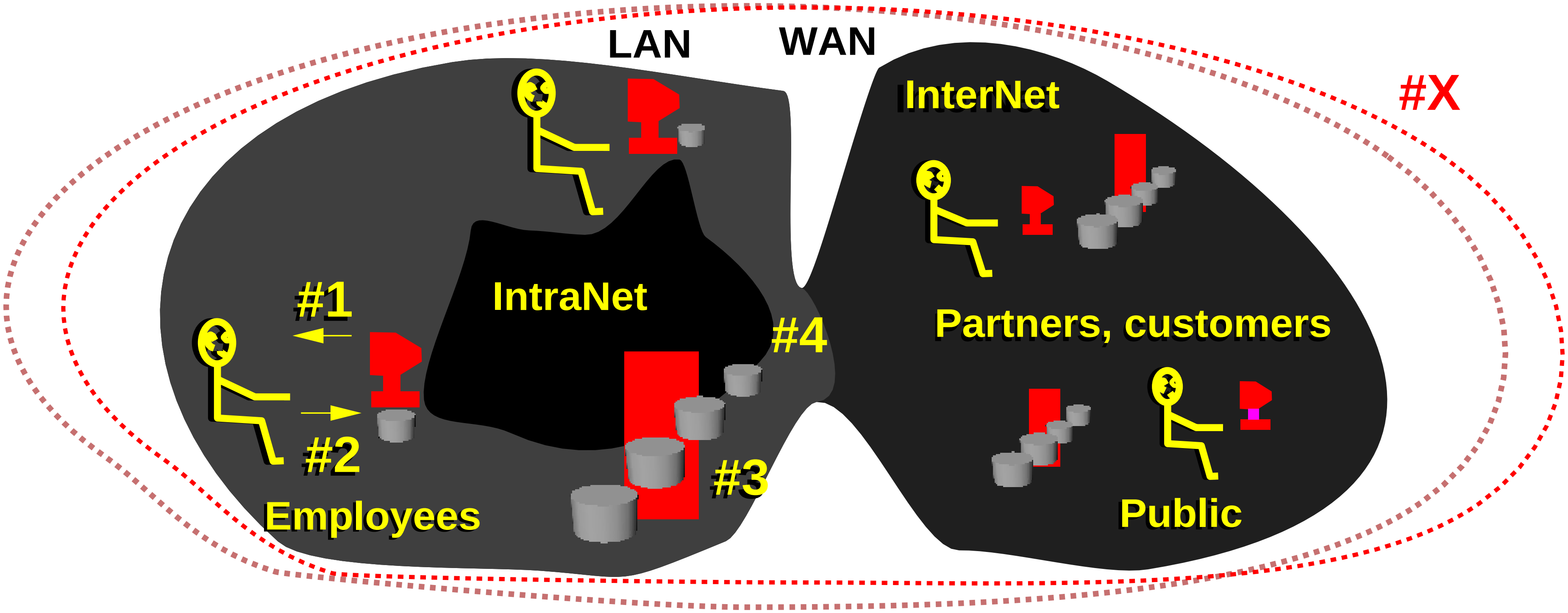
than in underlying *infrastructure:*

bandwidths, addressability & naming,
scalability of interconnect,
operating systems, file systems, backup ...

Symptoms: bottlenecks, odd limits, workarounds, instability, unpredictability, nonlinear surprise, over–frequent releases, multiple versions, hardware obsolete before depreciated

2. In organizations that grow quickly, stress on management and support infrastructure.

Environment:
4*X Data Problems



#1 Have data, cannot find & understand it	insight <− data
#2 Cannot create data from outside	creativity −> data
#3 Cannot have/process data, system limits	(data)
Server always needs (30%?) headroom	power
#4 Have the data, but in wrong place/form	data <−> data
Internal interconnect; network; firewalls	unleash

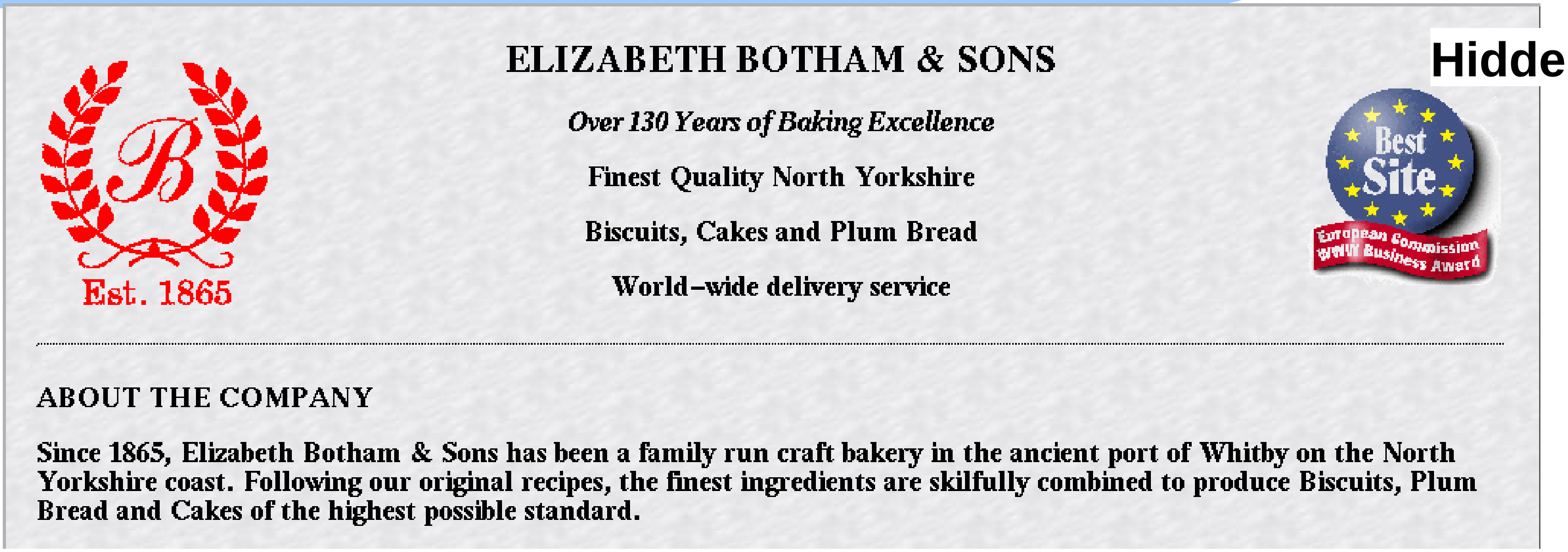
#X Rapid change, surprise amplify all 4 DATA problems

Data distribution more troublesome than CPU distribution

http://www.botham.co.uk

sgi

Hidden flag



ELIZABETH BOTHAM & SONS

Over 130 Years of Baking Excellence

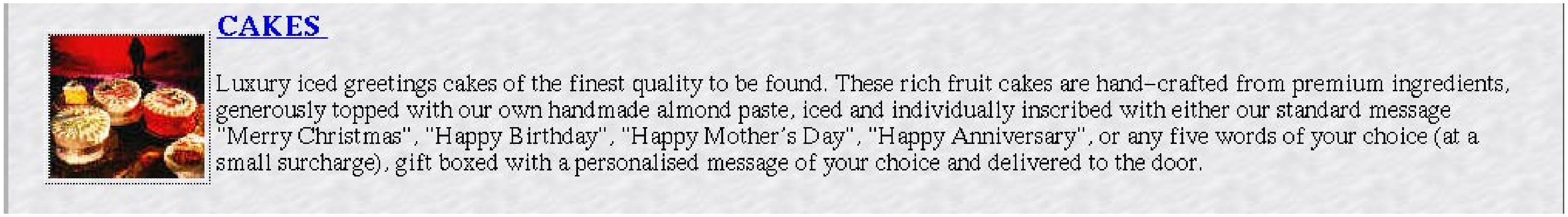
Finest Quality North Yorkshire

Biscuits, Cakes and Plum Bread

World–wide delivery service

ABOUT THE COMPANY

Since 1865, Elizabeth Botham & Sons has been a family run craft bakery in the ancient port of Whitby on the North Yorkshire coast. Following our original recipes, the finest ingredients are skilfully combined to produce Biscuits, Plum Bread and Cakes of the highest possible standard.



CAKES

Luxury iced greetings cakes of the finest quality to be found. These rich fruit cakes are hand–crafted from premium ingredients, generously topped with our own handmade almond paste, iced and individually inscribed with either our standard message "Merry Christmas", "Happy Birthday", "Happy Mother's Day", "Happy Anniversary", or any five words of your choice (at a small surcharge), gift boxed with a personalised message of your choice and delivered to the door.

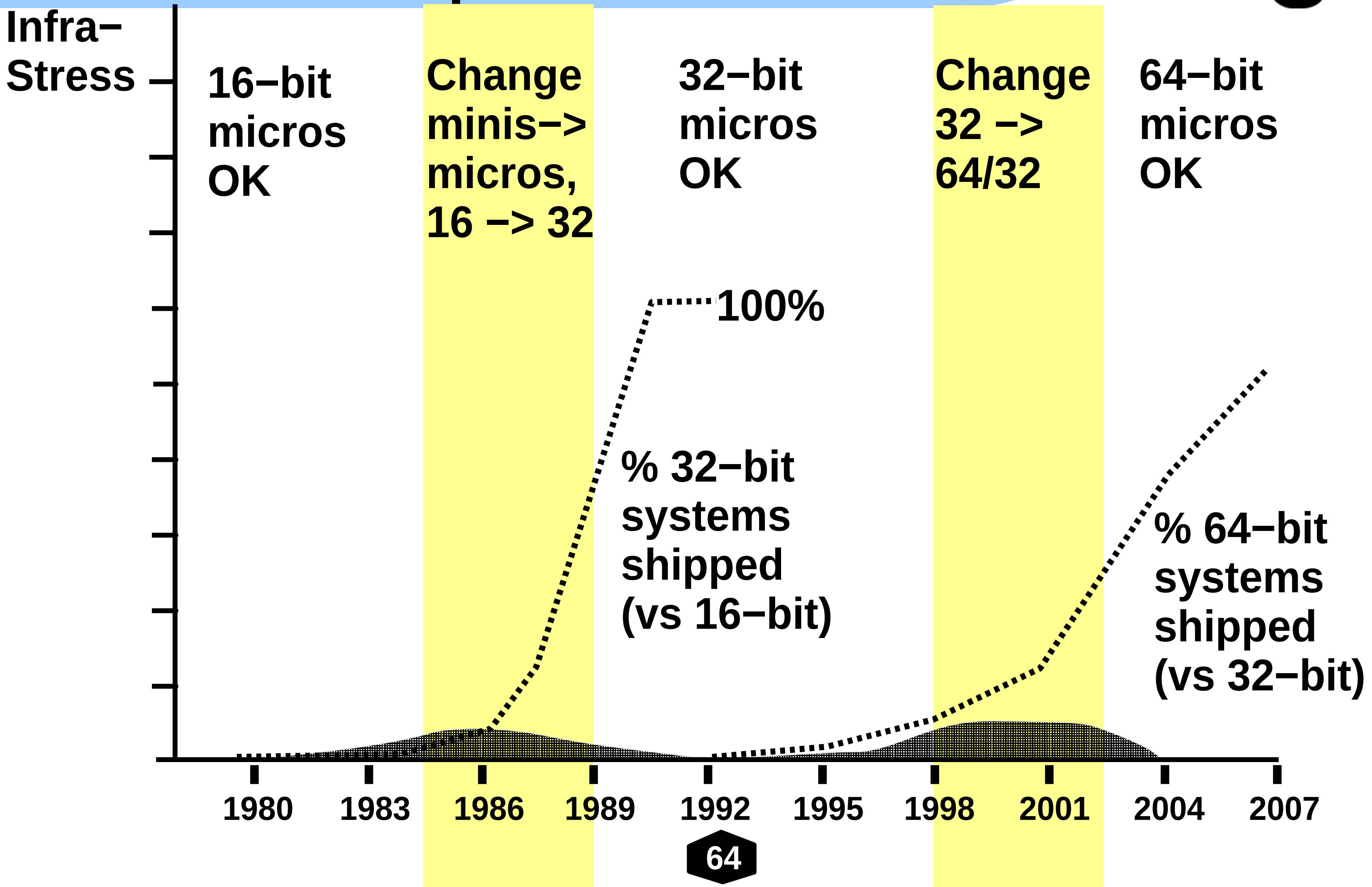
Family bakery in Yorkshire + Website
=> suddenly begin selling outside UK.

Predict this?

No ... just predict change & surprise.
But, some technology predictions easier...

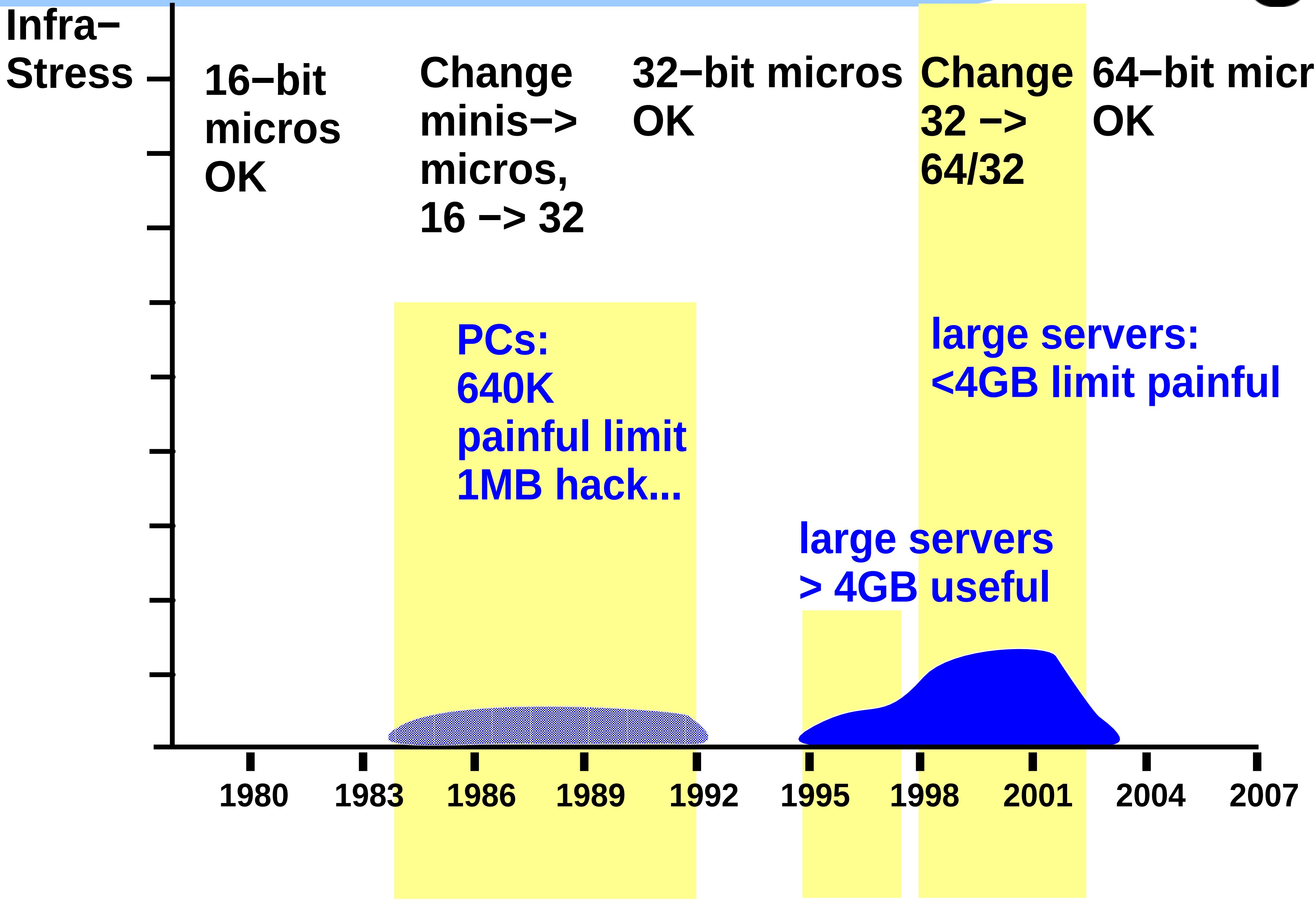
1. CPUs

CMOS Microprocessors



1st 64–bit micro (MIPS R4000)

2. Big Memory & Micros



16–bit micros OK

Change minis–> micros, 16 –> 32

32–bit micros OK

Change 32 –> 64/32

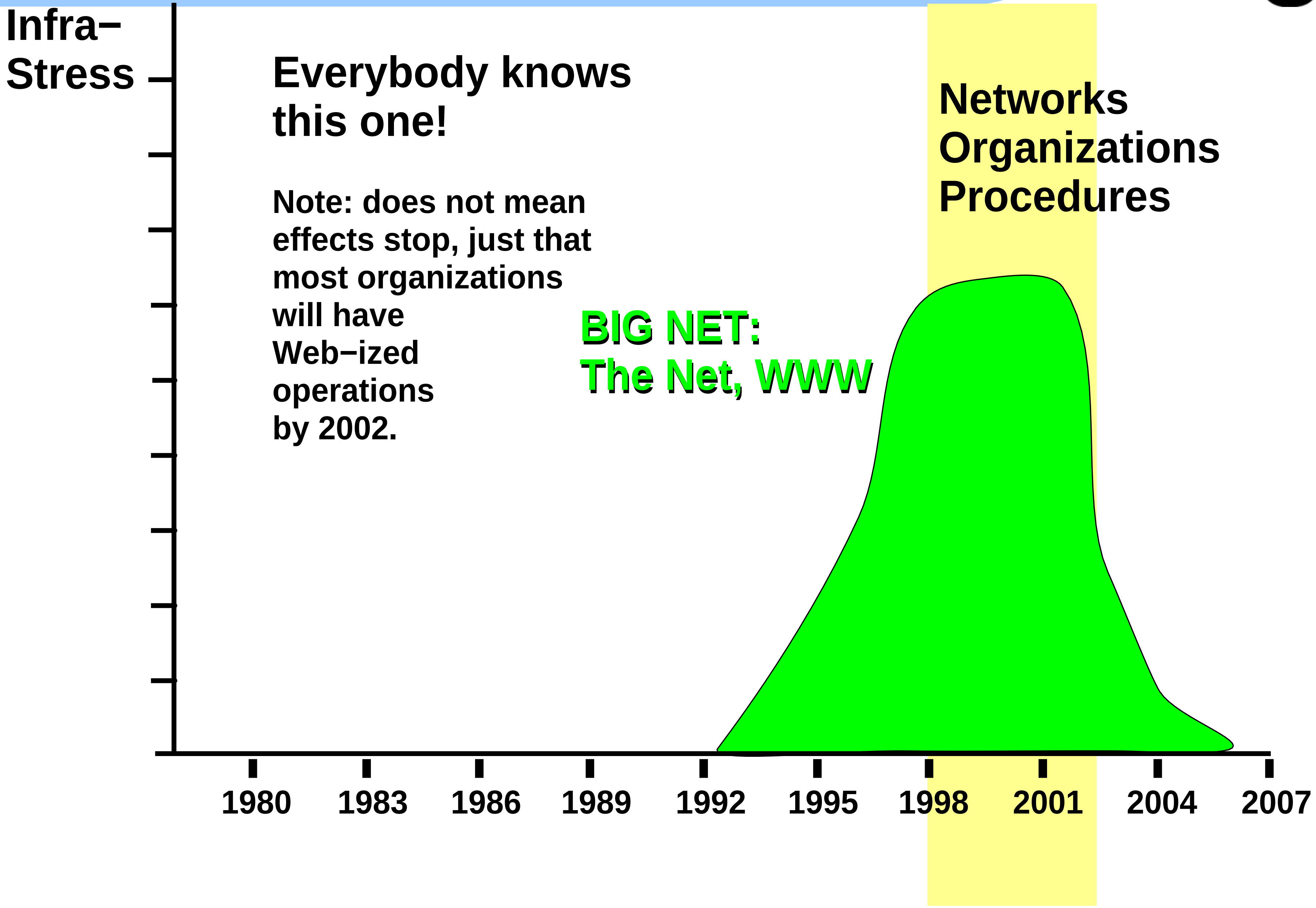
64–bit micros OK

PCs: 640K painful limit 1MB hack...

large servers: <4GB limit painful

large servers > 4GB useful

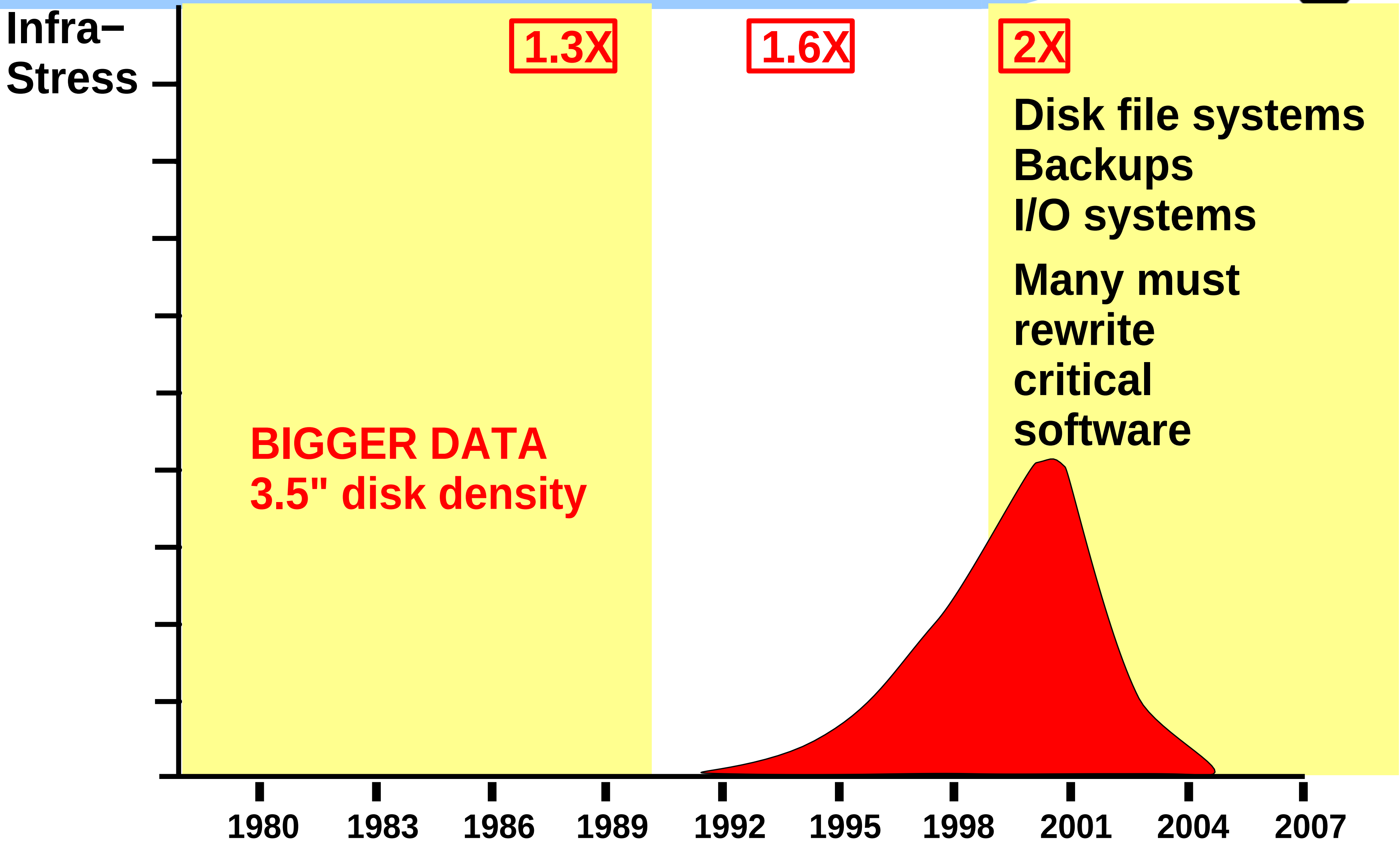
3. Big Net



Everybody knows this one!

Note: does not mean effects stop, just that most organizations will have Web−ized operations by 2002.

4. Bigger (Disk) Data



BIGGER DATA
3.5" disk density

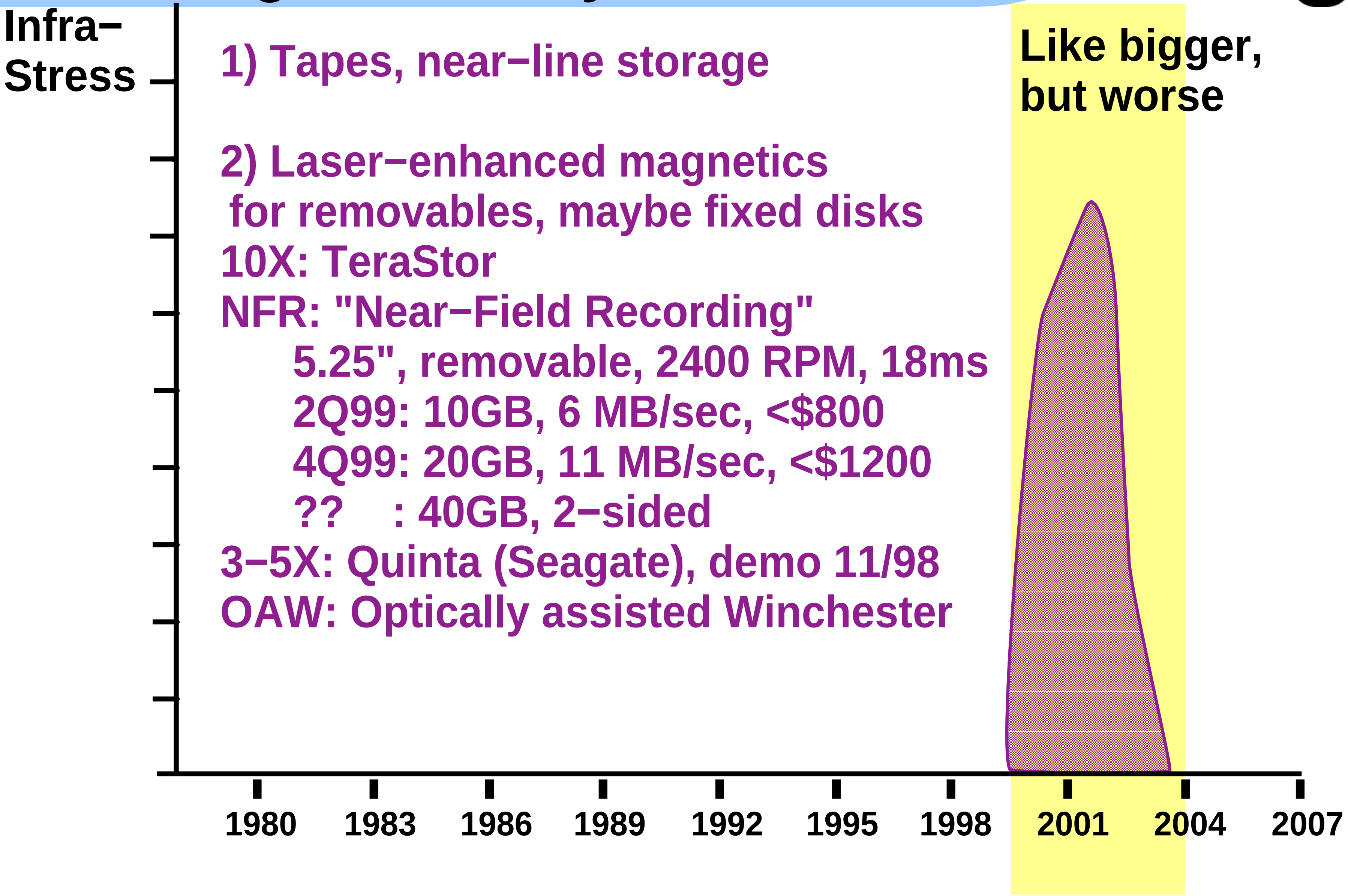
Disk file systems
Backups
I/O systems

Many must
rewrite
critical
software

http://www.quantum.com/src/history, http://www.disktrend.com
http://www.ibm.com/storage/microdrive: 340MB Microdrive, 1999. 1.7"x1.4"x.19"

5. HUGE Data (Maybe)

Storage Hierarchy



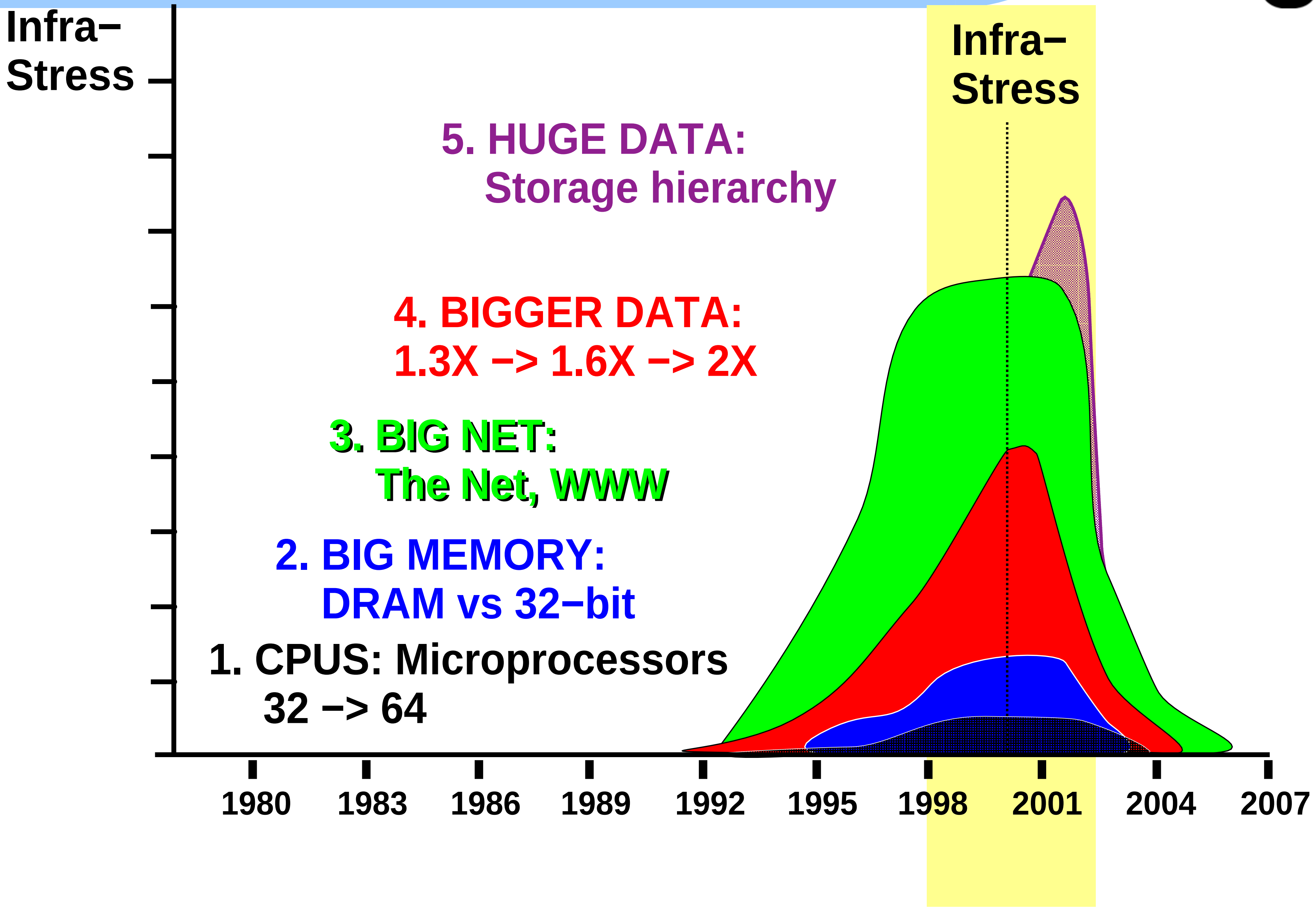
Like bigger,
but worse

1) Tapes, near−line storage

2) Laser−enhanced magnetics
for removables, maybe fixed disks
10X: TeraStor
NFR: "Near−Field Recording"
5.25", removable, 2400 RPM, 18ms
2Q99: 10GB, 6 MB/sec, <$800
4Q99: 20GB, 11 MB/sec, <$1200
?? : 40GB, 2−sided
3−5X: Quinta (Seagate), demo 11/98
OAW: Optically assisted Winchester

~1999: Laser=enhanced magnetic disks (removable)
http://www.quinta.com, http://www.terastor.com

InfraStress Addup



5. HUGE DATA:
   Storage hierarchy

4. BIGGER DATA:
   1.3X –> 1.6X –> 2X

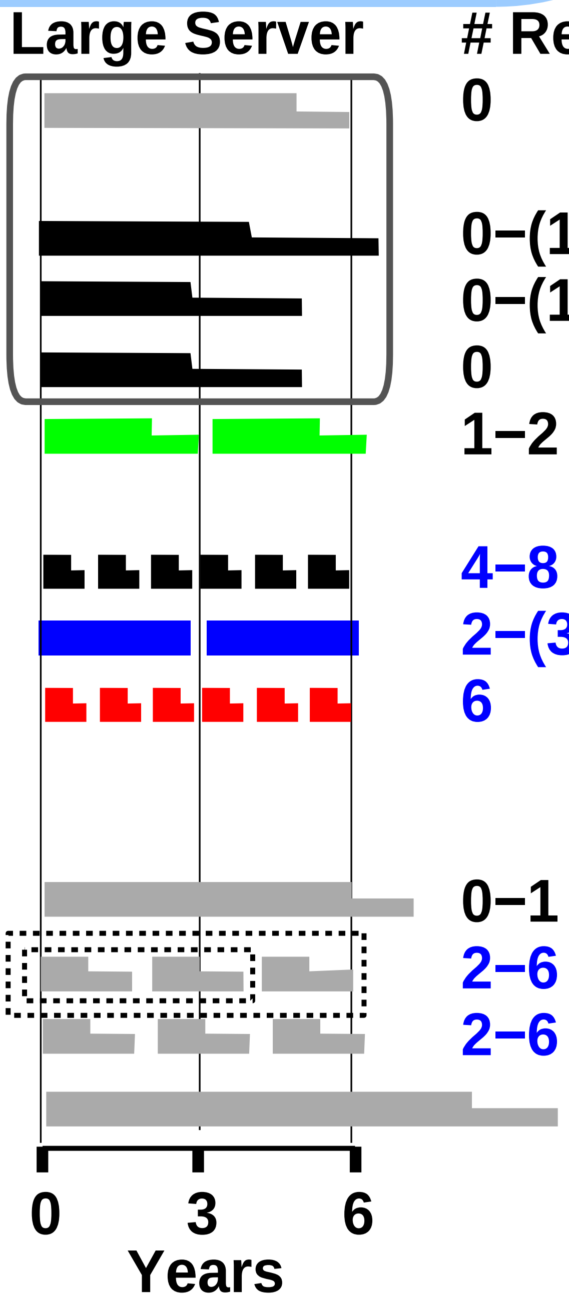
3. BIG NET:
   The Net, WWW

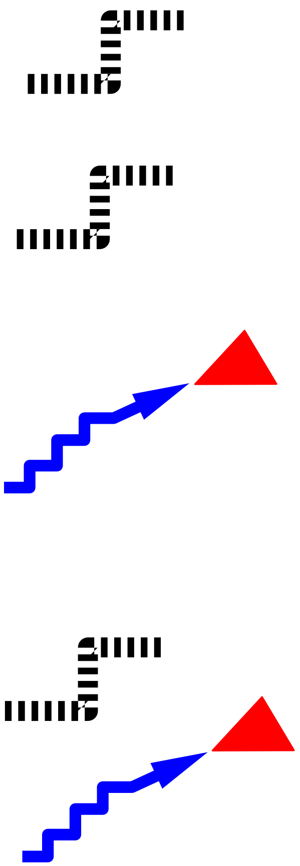
2. BIG MEMORY:
   DRAM vs 32–bit

1. CPUS: Microprocessors
   32 –> 64

Technology Change Rates
Example: Large Server*

	Years	# Revisions in 6 years
H/W chassis	4..6	0
Interconnects		
I/O bus (PCI...)	4−6+	0−(1)
CPU==mem	3−5	0−(1)
Backplane	3−5	0
Network	varies	1−2
Subsystems		
CPU MHz	.75−1.5	4−8
4X DRAM	3	2−(3)
Disks	1	6
Graphics	1.5−2.5	
Software		
File system	8−10	0−1
OS release	1−2	2−6
App release	1−2	2−6
Data	forever	
Media	not long	





*Desktops &
other access devices
cycle faster, maybe

Technology Trends

Capacities − Great News

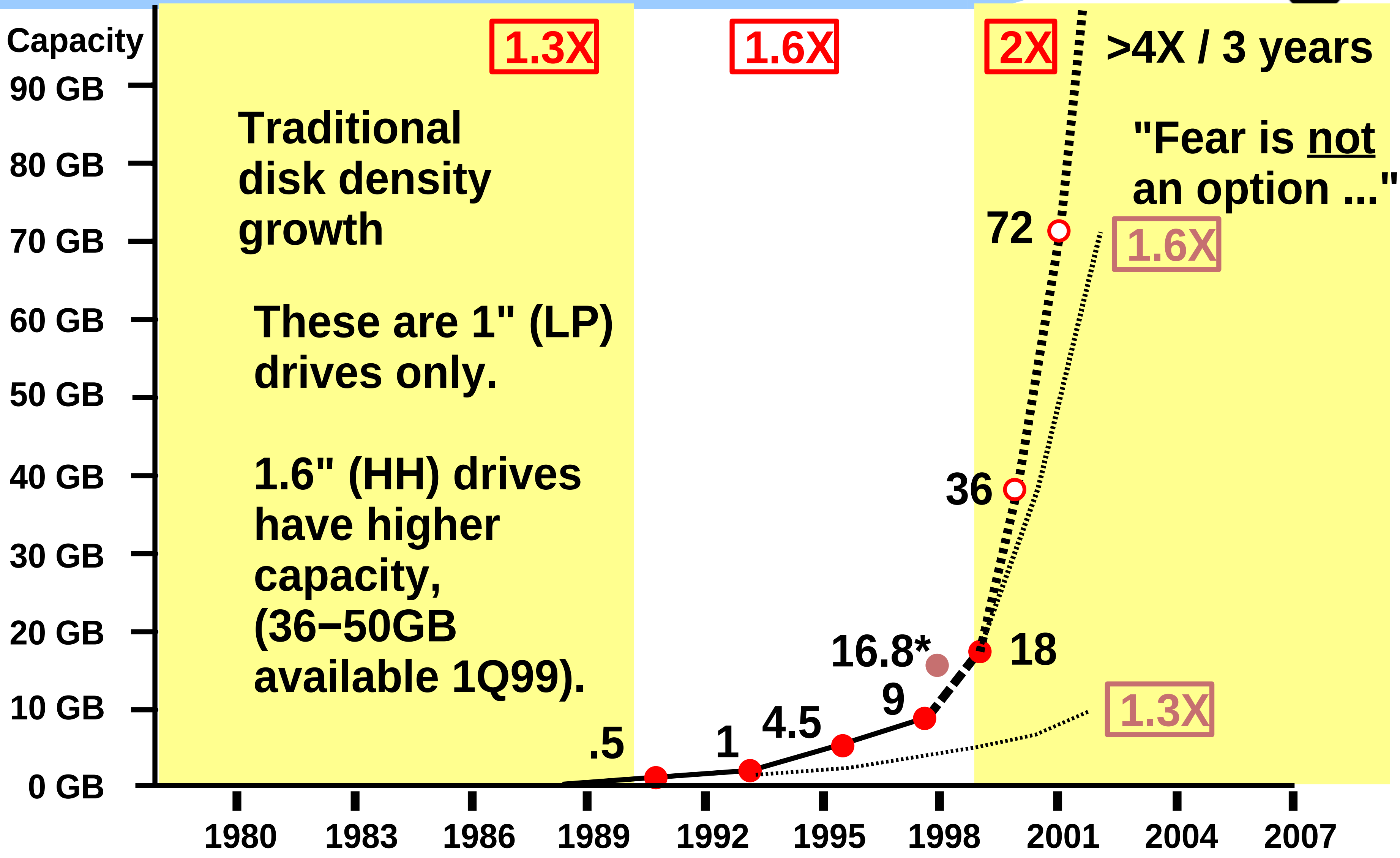
Latencies − Not−so−great News

Bandwidths − InfraStress

Interactions − Surprises

Tradeoffs − keep changing

1"x 3.5" Disk Capacity



>4X / 3 years

Traditional disk density growth

These are 1" (LP) drives only.

1.6" (HH) drives have higher capacity, (36−50GB available 1Q99).

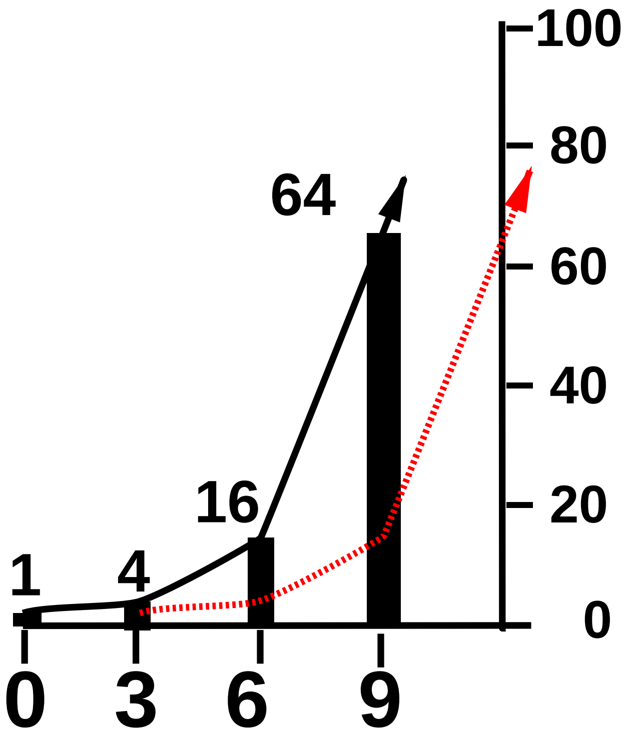
"Fear is not an option ..."

"Disks are binary devices ... new and full"

*IBM Desktap 16GP, Giant Magnetoresistive heads (GMR), 4Q97.

Log–scale charts ahead

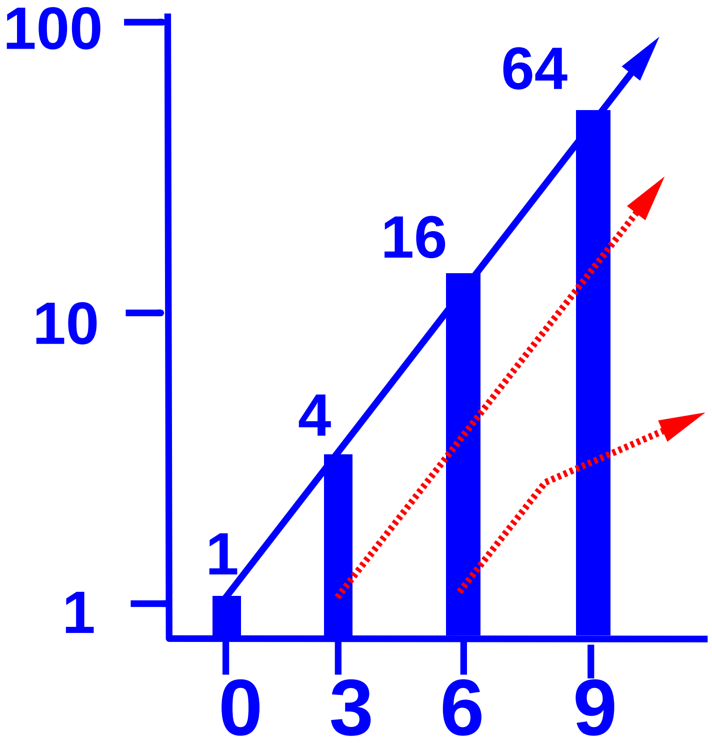
Linear scale



==>

Logarithmic scale

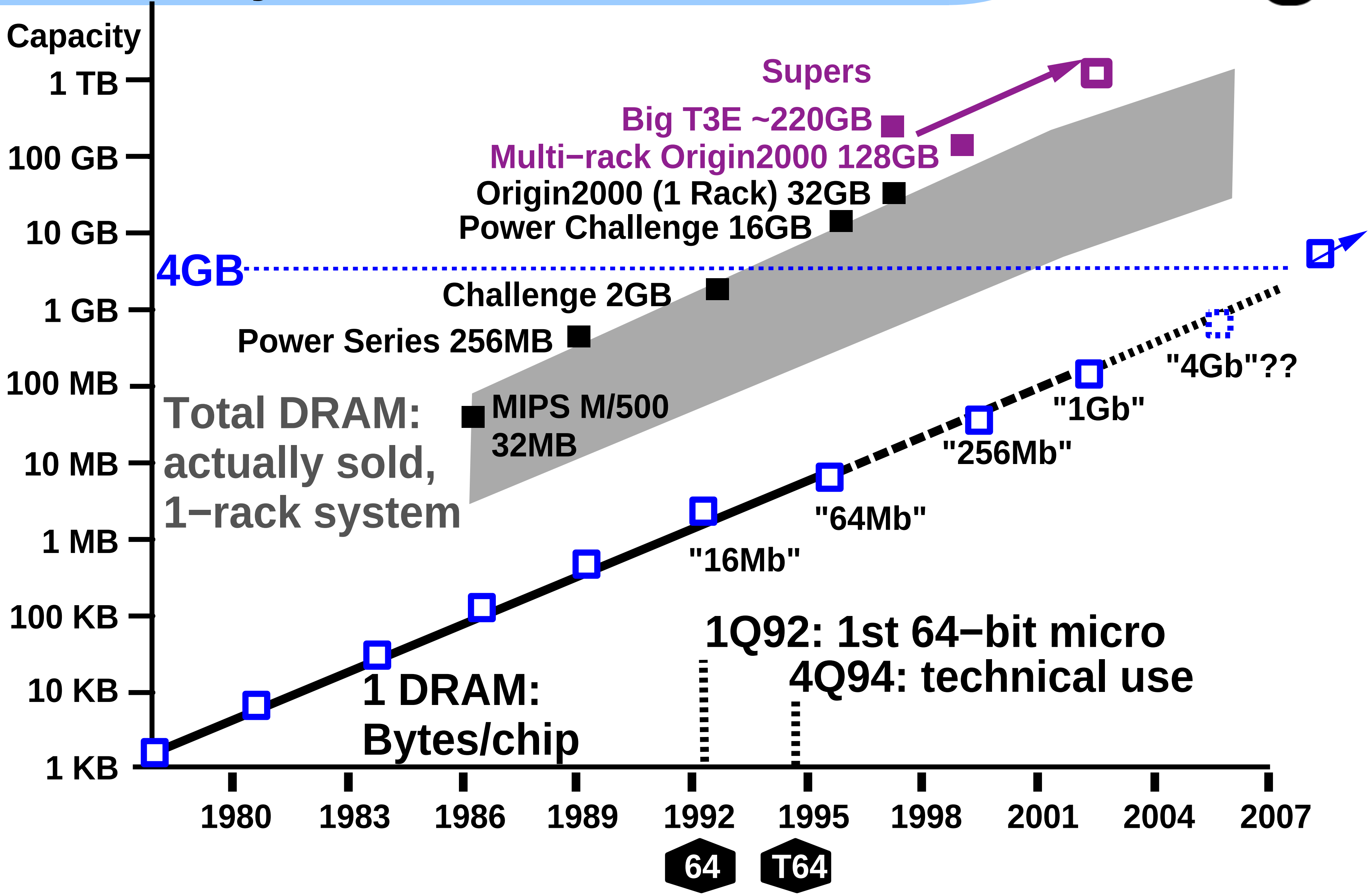
Huge differences do not *look* so big at top



Parallel = same ratio
Inflection points clear

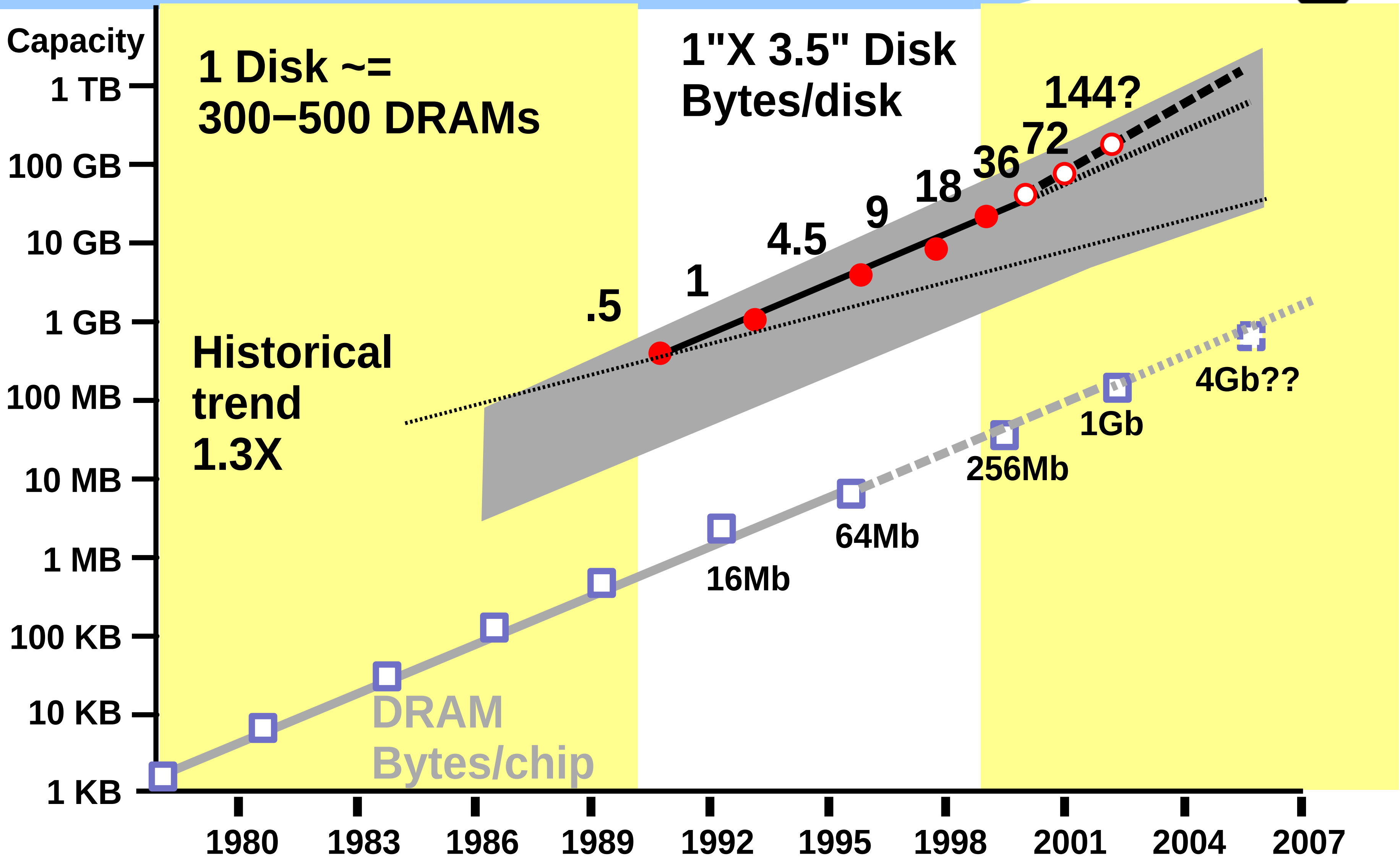
DRAM Capacity: 1.6X CAGR

4X / 3 years



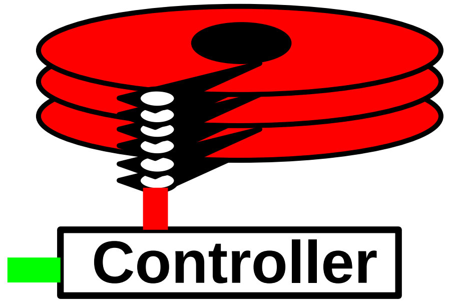
See: John R. Mashey, "64−bit Computing", BYTE, September 1991, 135−141.

Disk Capacity:
1.3X −> 1.6X −> 2X



See: John R. Mashey, Darryl Ramm, "Databases on RISC: still The Future", UNIX Review, September 1996, 47−54.

3.5" Disk Review



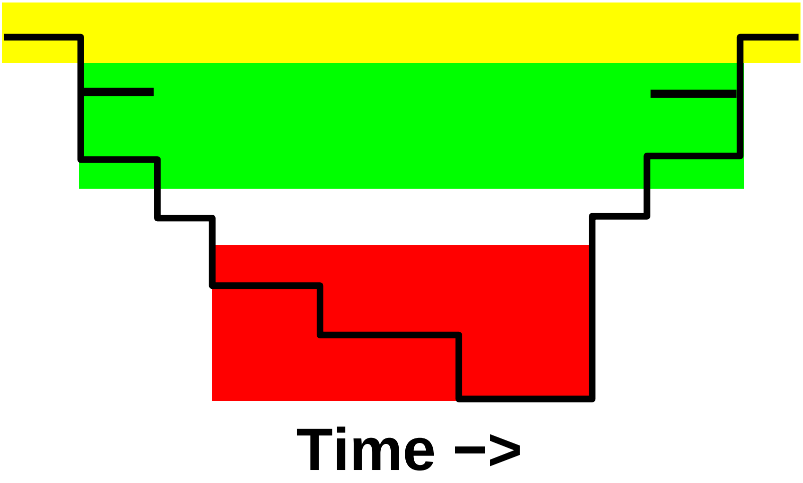
- Height (1" or 1.6") X (4" X 5.75")
- Capacity (1MB = 1,000,000 B)
- Seek Times (msecs)
  - Track−to−track (Read/Write)
  - Average (Read/Write)
    - Typical < Average (OS & controllers)
  - Maximum (Read/Write)
- Rotational latency (msecs)
  - Average Latency = .5 * rev = 30000/RPM
- Bandwidths (MB/sec)
  - Internal Formatted Transfer
  - ZBR range
  - External Rate (Bus)
- Density (Gbit/sq inch)

See:http://www.quantum.com/src/basic_resources
See "Disk Performance Background for Tables/Graphs", SGI internal, Radek Aster, Jeremey Higdon, Carl Rigg, June 27, 1997.

3.5" Disk Review

- Capacity/drive ~ # platters (varies)
- Capacity/platter ~ areal density
- Bandwidth ~ RPM * Linear density
- Seek time ... improves slowly
- Combine several drives onto one: take care, may lose seeks/second
- IOPS vs MB/s applications

System (OS)
I/O Bus (~PCI)
Peripheral Connect (~SCSI)
Embedded Disk Controller
Disk Seek
Rotate
Read



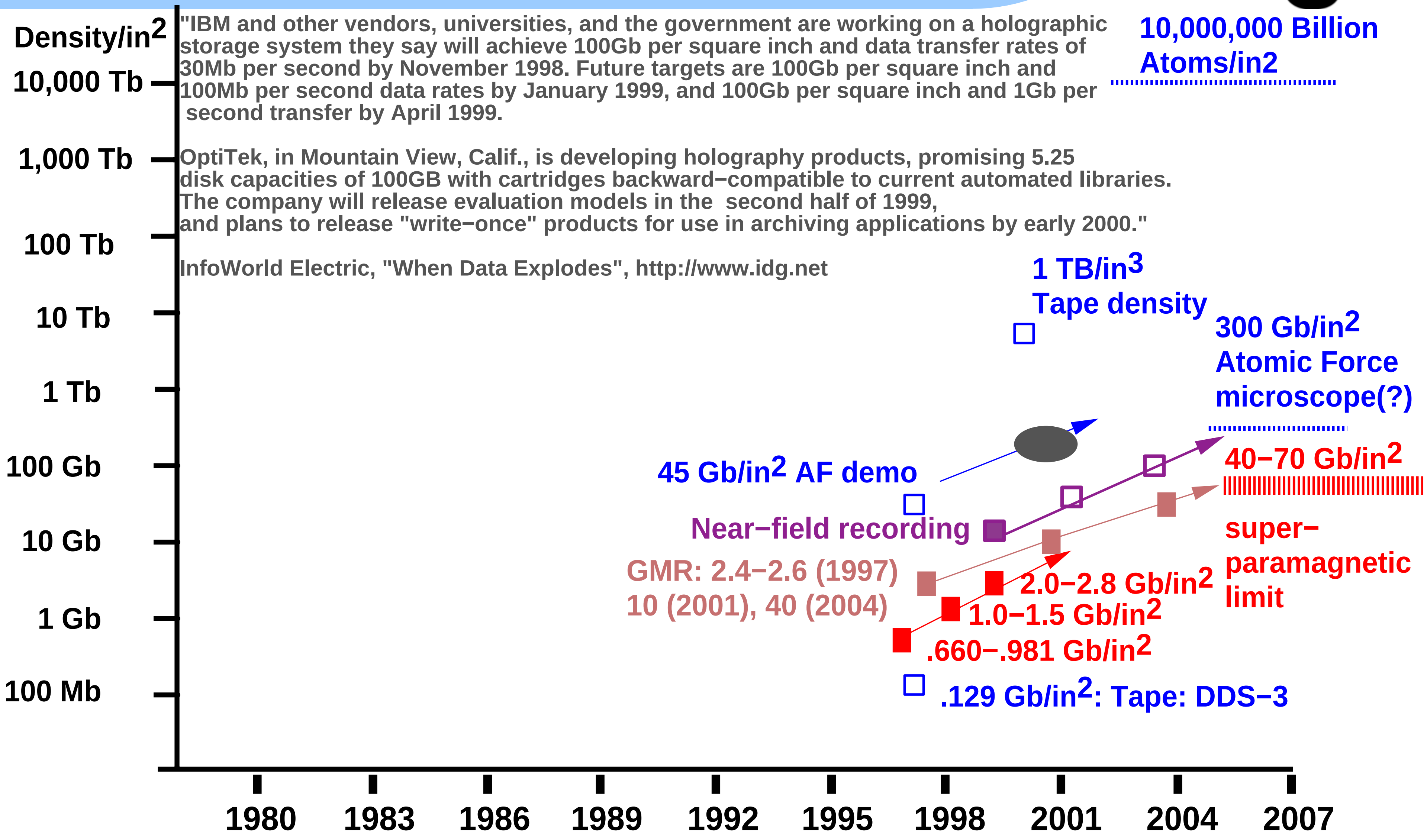
Common Disk Types

1. By capacity
   A. Large (1.6" x 3.5", HH) ~8−10 platters
   B. Medium (1" X 3.5", LP), ~4−5 platters
   C. "Depopulated", 1 platter
   D. Smaller platters ...
   E. "Microdrive", 1 small platter

2. By target
   – High−performance (B: high RPM)
   – High−capacity (A)
   – By IOPs (multiples of C & D)
   – By cost [ATA, IDE versions of A, B, C]
   – By physical size (mobile, consumer)Bad

Huge disks => long backup times
Good for archive−like applications

Storage Densities



"IBM and other vendors, universities, and the government are working on a holographic storage system they say will achieve 100Gb per square inch and data transfer rates of 30Mb per second by November 1998. Future targets are 100Gb per square inch and 100Mb per second data rates by January 1999, and 100Gb per square inch and 1Gb per second transfer by April 1999.

OptiTek, in Mountain View, Calif., is developing holography products, promising 5.25 disk capacities of 100GB with cartridges backward−compatible to current automated libraries. The company will release evaluation models in the second half of 1999, and plans to release "write−once" products for use in archiving applications by early 2000."

InfoWorld Electric, "When Data Explodes", http://www.idg.net

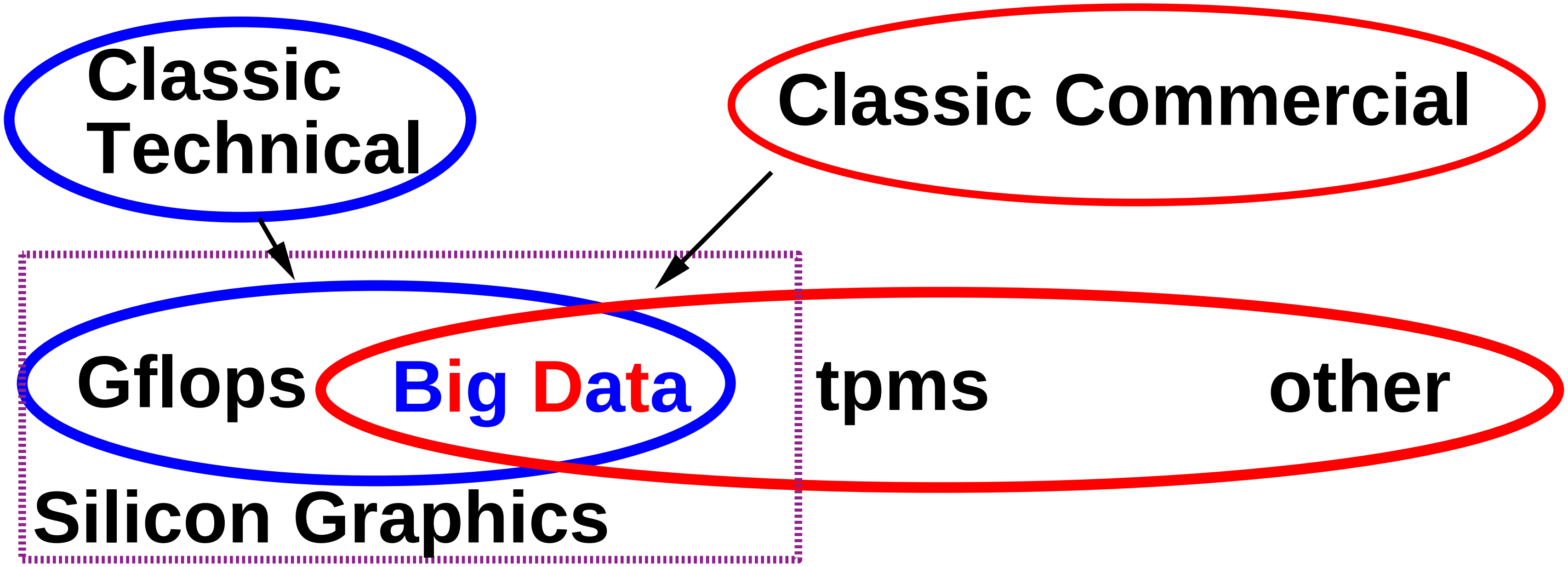
See: Merrit E. Jones, The MITRE Corp, "The Limits That Await Us", THIC Meeting April 23, 1997, Falls Church, Va.
See http://www.terastor.com on Near−field recording.

Disk Issues
Workloads Converge

"IOPS" – Transaction / seeks/second
Classic OLTP, small blocks

"MB/s" – Bandwidth (& backup!)
Classic technical, larger blocks

Some commercial now more like technical



Disk Issues – Implications

1. Huge capacity leap breaks old filesystems
   Hard limits (2GB, 8GB, etc) OR
   Algorithmic performance, scaling issues

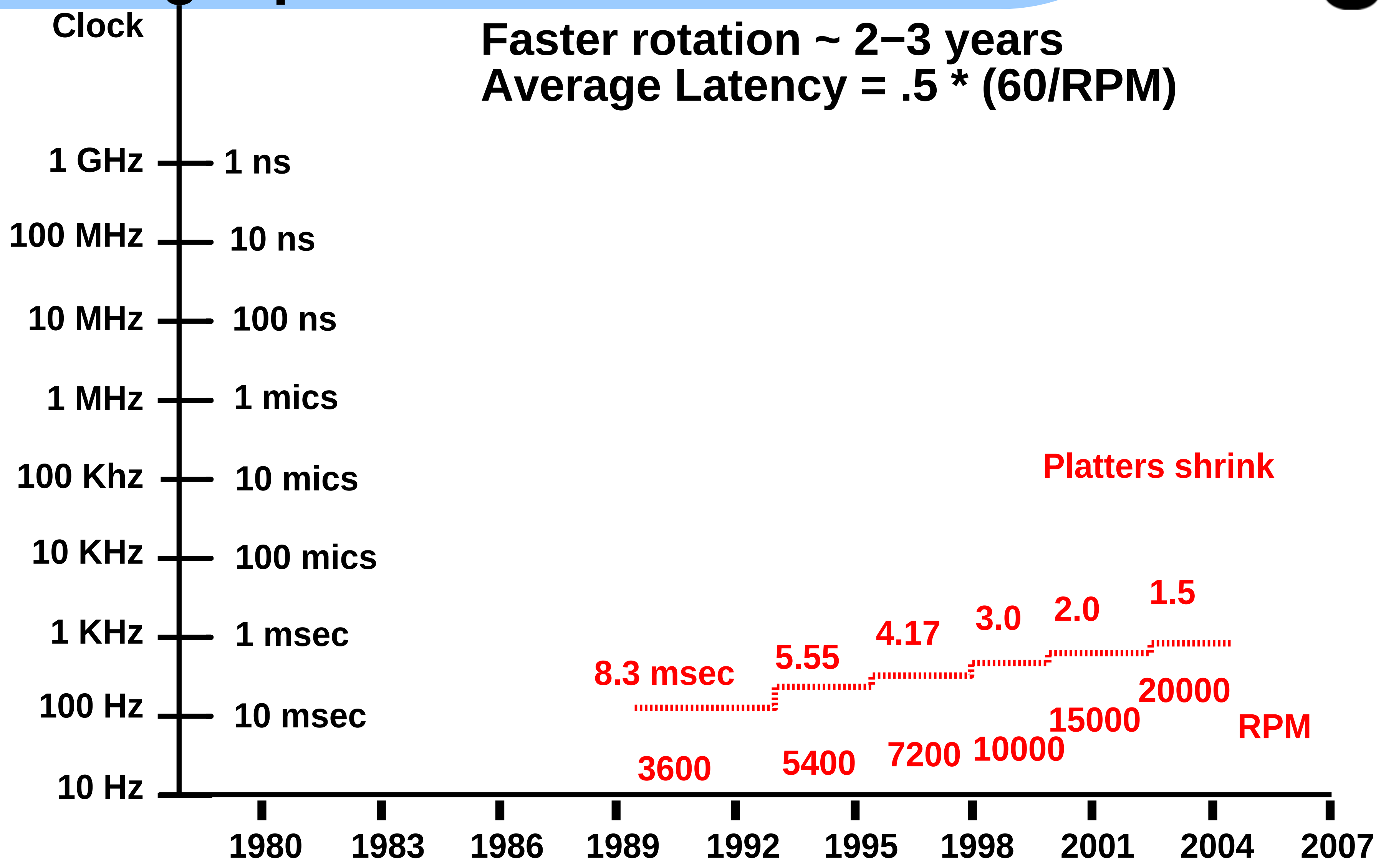
2. More memory, more bandwidth, everywhere
   Small disk blocks even less efficient
   => 64–bit addressing more useful
   => Big pages, map more pages, MMUs
   => More memory => more bandwidth
   => More interconnect bandwidth

3. BACKUP ...
   Must run many tapes, full–speed, parallel
   Sometimes use HSM, RAID, mirror
   New cartridge disks may be useful

Disk Rotational Latencies

High−performance − 1/2 Rotation

Faster rotation ~ 2−3 years
Average Latency = .5 * (60/RPM)



Money can buy bandwidth, but latency is forever.

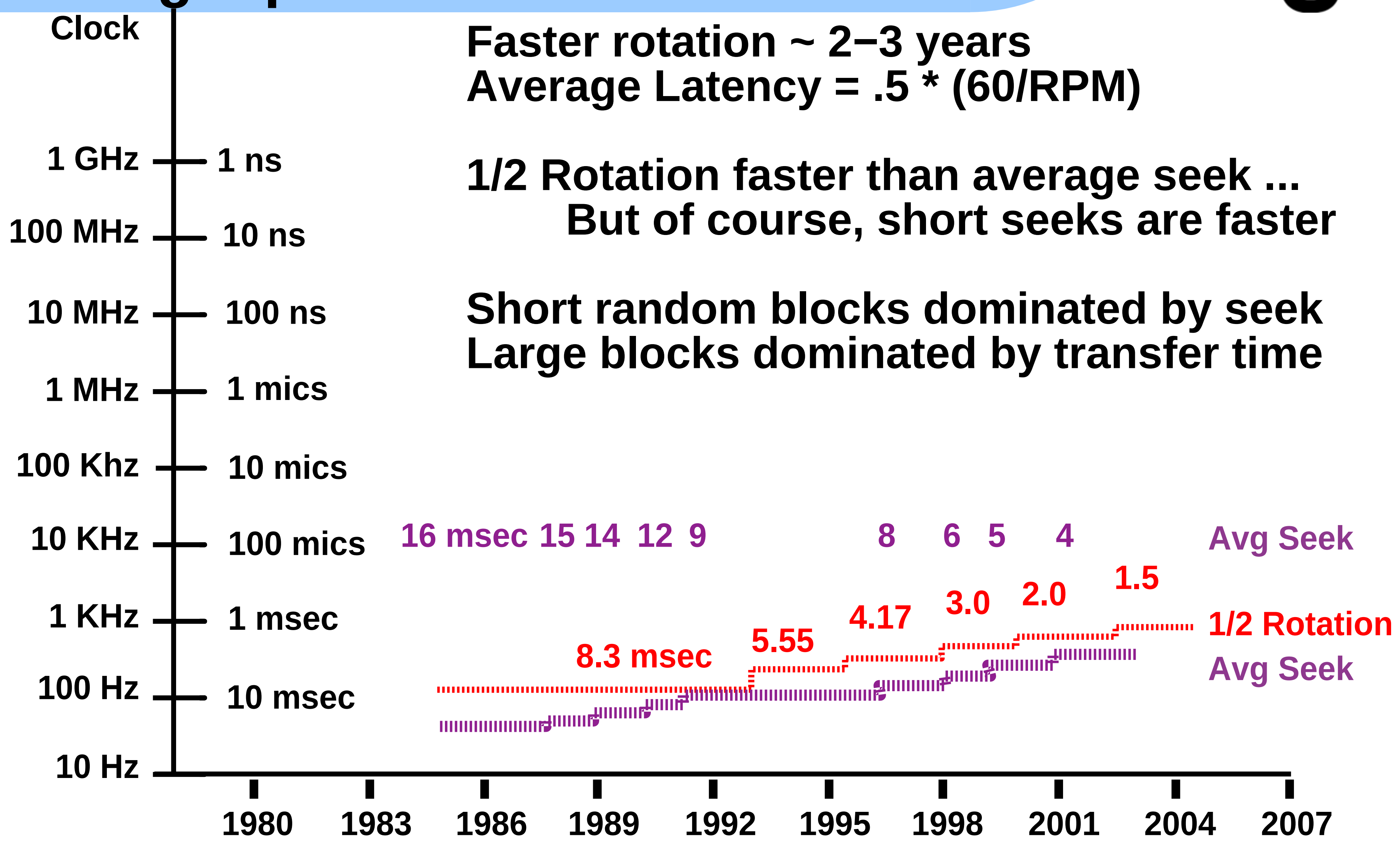
Disk Average Seek

High−performance disks

Faster rotation ~ 2−3 years
Average Latency = .5 * (60/RPM)

1/2 Rotation faster than average seek ...
But of course, short seeks are faster

Short random blocks dominated by seek
Large blocks dominated by transfer time



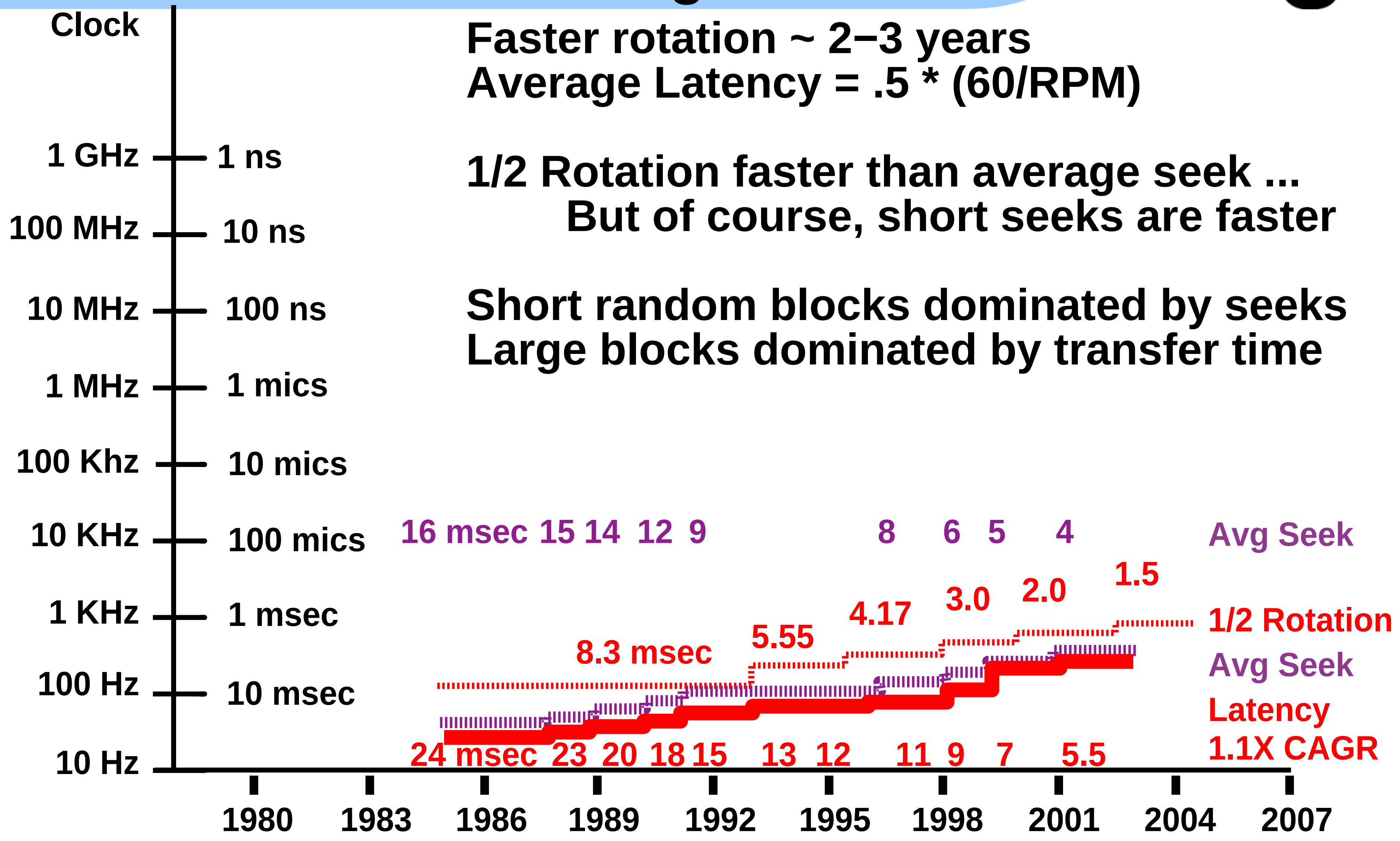
Disk Total Latencies

1/2 Rotation + Average Seek

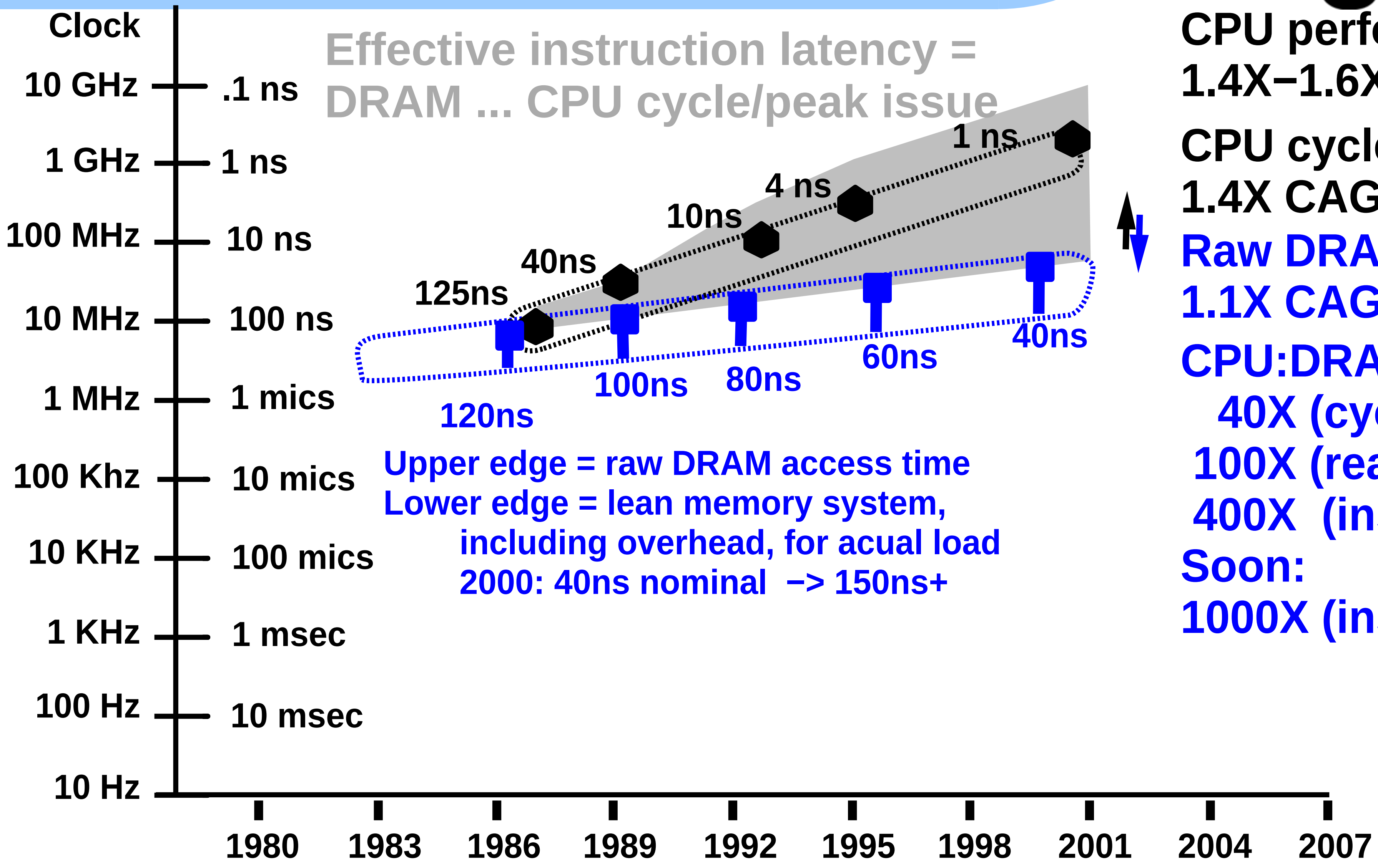
Faster rotation ~ 2−3 years
Average Latency = .5 * (60/RPM)

1/2 Rotation faster than average seek ...
But of course, short seeks are faster

Short random blocks dominated by seeks
Large blocks dominated by transfer time



CPU Latency, Performance



Effective instruction latency = DRAM ... CPU cycle/peak issue

Upper edge = raw DRAM access time
Lower edge = lean memory system,
including overhead, for acual load
2000: 40ns nominal −> 150ns+

CPU perform
1.4X−1.6X

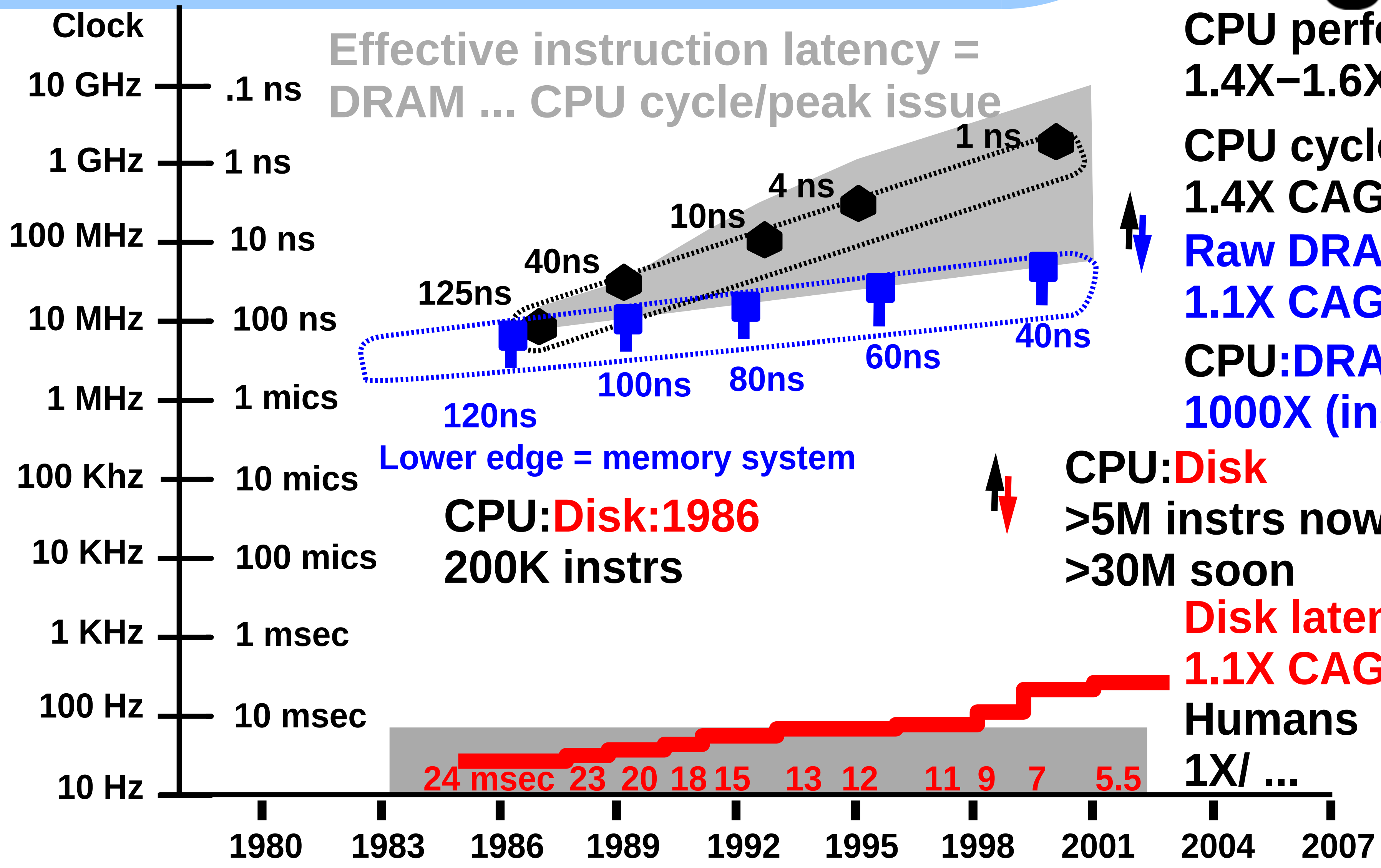
CPU cycle
1.4X CAGR

Raw DRAM
1.1X CAGR

CPU:DRAM:
40X (cycle)
100X (real)
400X (instrs)

Soon:
1000X (instrs)

Latency & Performance



Effective instruction latency = DRAM ... CPU cycle/peak issue

Lower edge = memory system

CPU:Disk:1986
200K instrs

CPU perform
1.4X−1.6X

CPU cycle
1.4X CAGR

Raw DRAM
1.1X CAGR

CPU:DRAM
1000X (insts)

CPU:Disk
>5M instrs now
>30M soon

Disk latency
1.1X CAGR

Humans
1X/ ...

Latencies – Implications

1. CPU <–> DRAM <–> disk
   Latency ratios already bad, getting worse.
   "Money can buy bandwidth, but latency is forever."

   ==> More latency tolerance in CPUs
   ==> Trade (bandwidth, memory, CPU,
   PROGRAMMING) for latency
   ==> Already worth 1M instruction
   to avoid a disk I/O

2. RDBMS huge buffer areas for indices,
   small tables, to avoid latency

3. Networks: be alert for latency issues

Input/Output: A Sad History

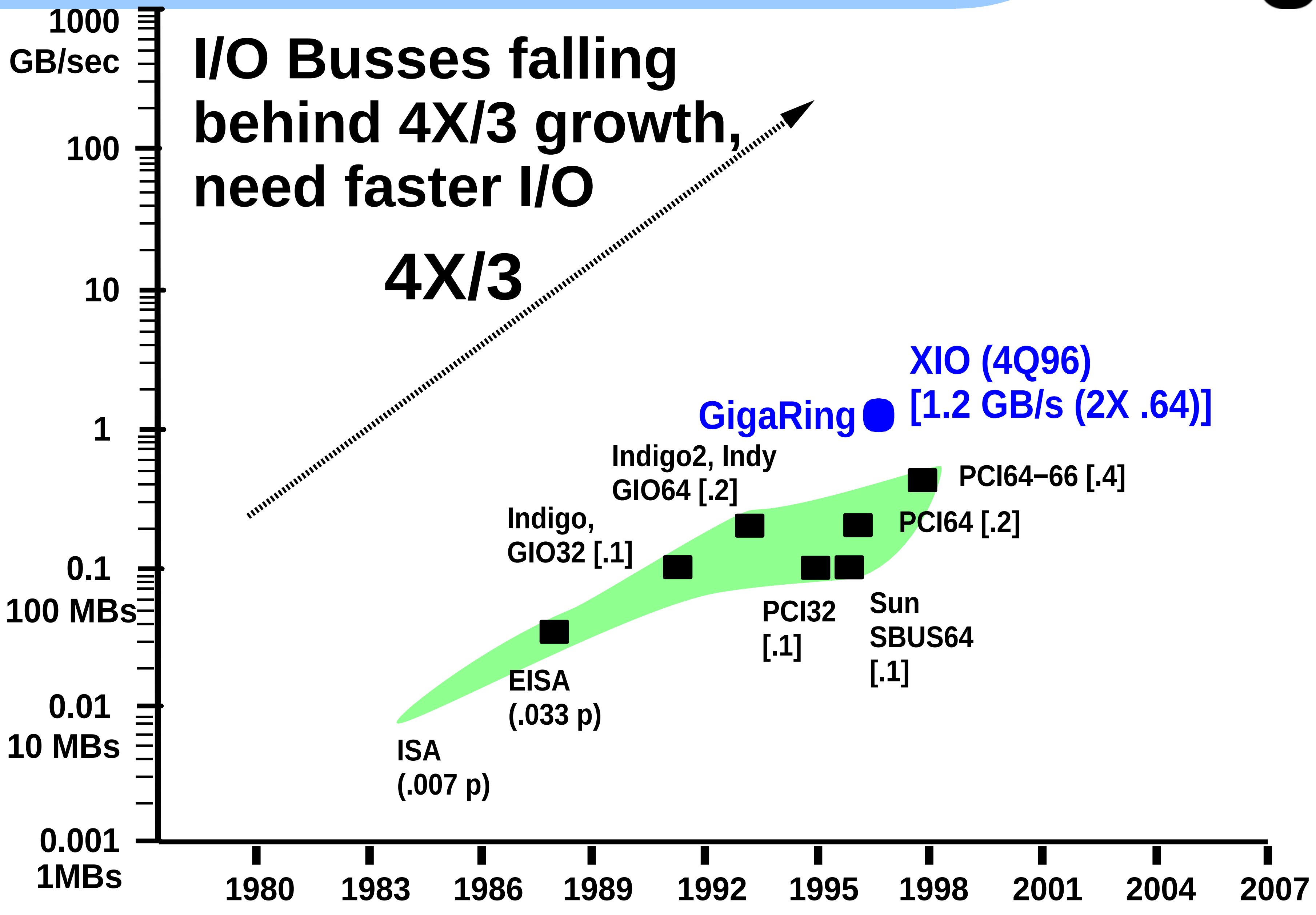
"I/O certainly has been lagging in the last decade."
Seymour Cray
Public Lecture (1976)

"Also, I/O needs a lot of work."
David Kuck
Keynote Address, 15th Annual Symposium
on Computer Architecture (1988)

"Input/output has been the orphan of computer architecture ... I/O’s revenge is at hand"
David A. Patterson, John. L. Hennessy
Computer Architecture: A Quantitative Approach,
2nd Ed (1996), Morgan Kaufmann.

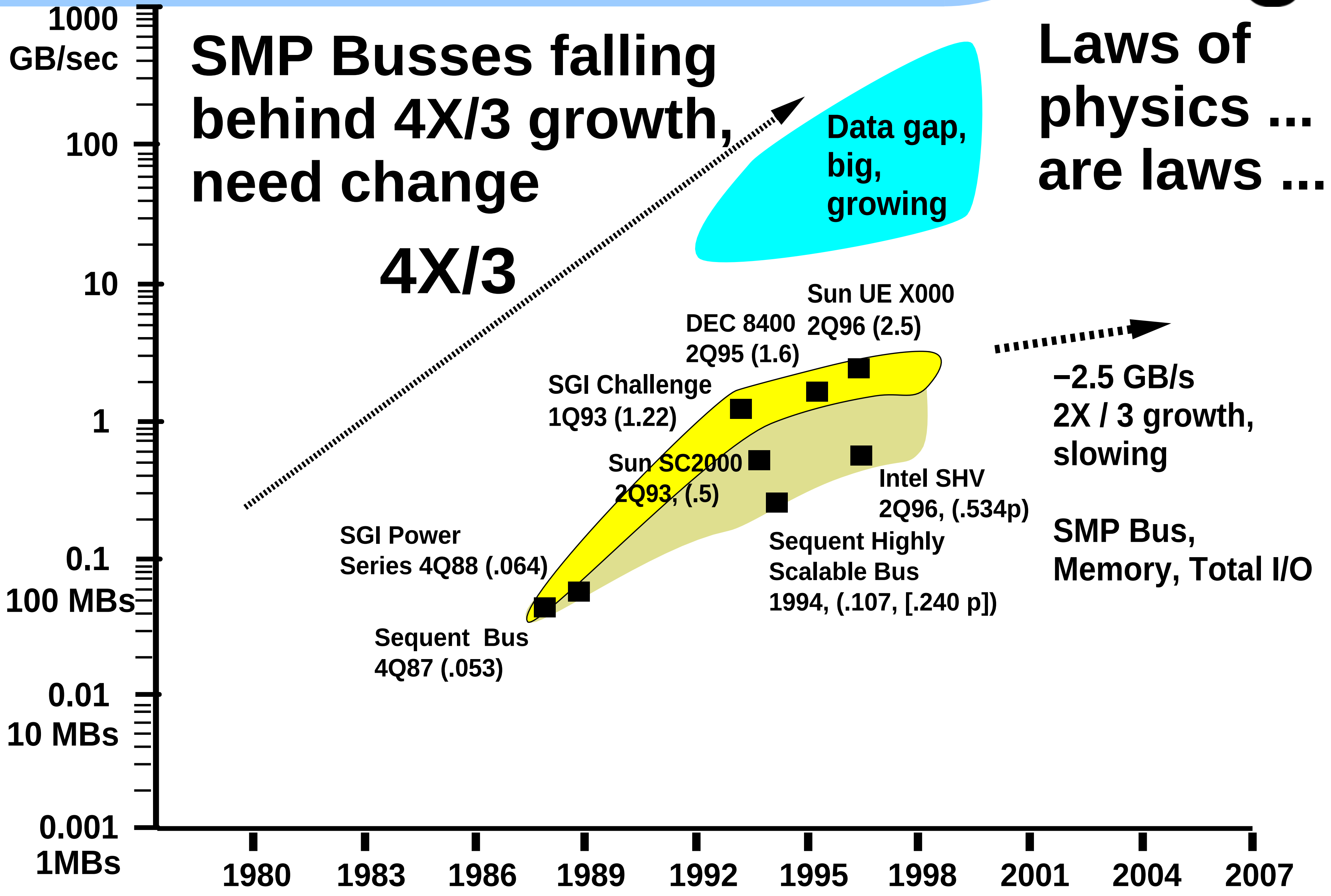
.

I/O Single−Channel Bandwidth



I/O Busses falling behind 4X/3 growth, need faster I/O

Bus–Based SMP
Bandwidth Wall



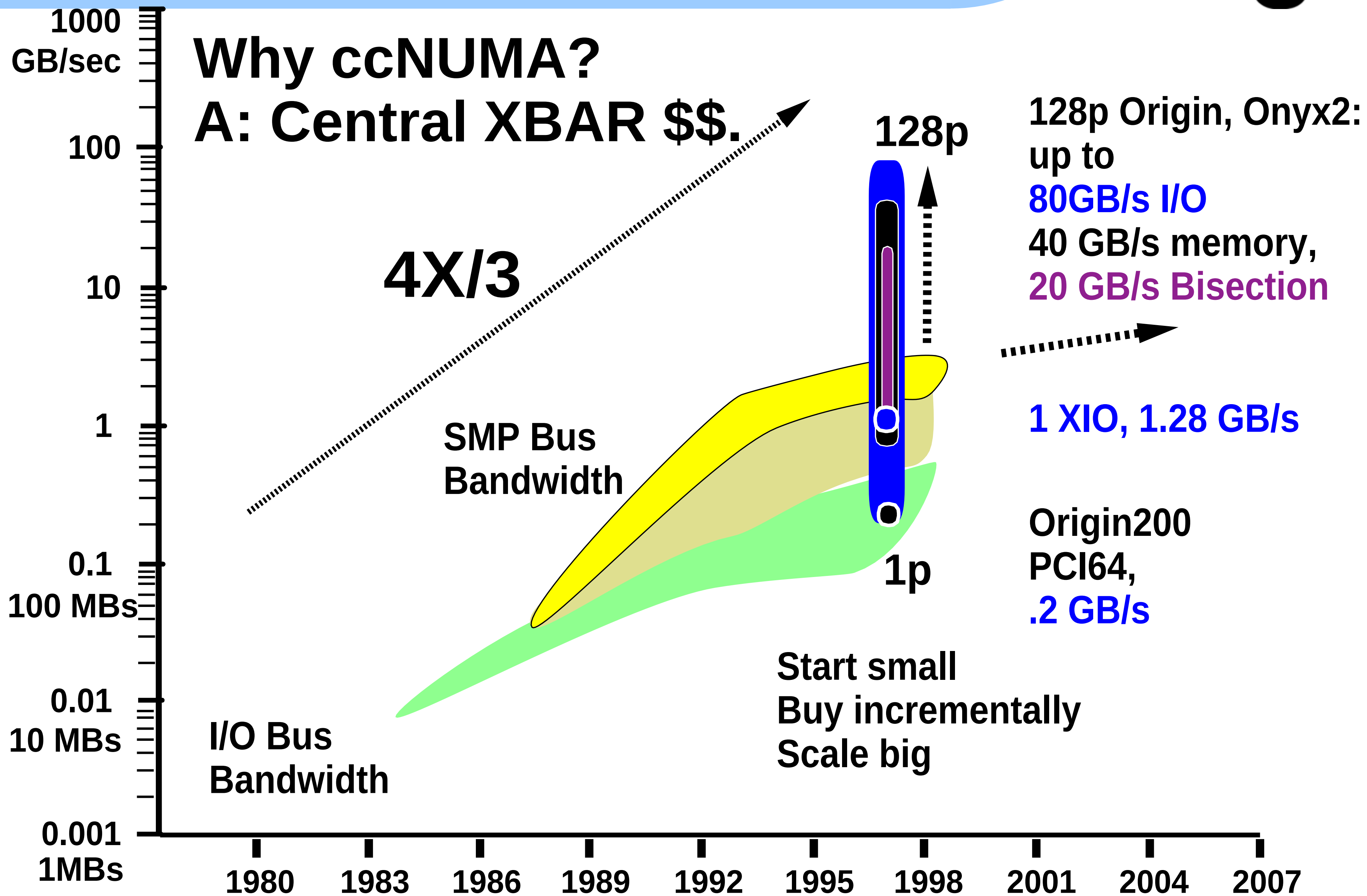
SMP Busses falling behind 4X/3 growth, need change

Laws of physics ... are laws ...

−2.5 GB/s
2X / 3 growth,
slowing

SMP Bus,
Memory, Total I/O

Bandwidths (ccNUMA, XBAR)



Why ccNUMA?
A: Central XBAR $$.

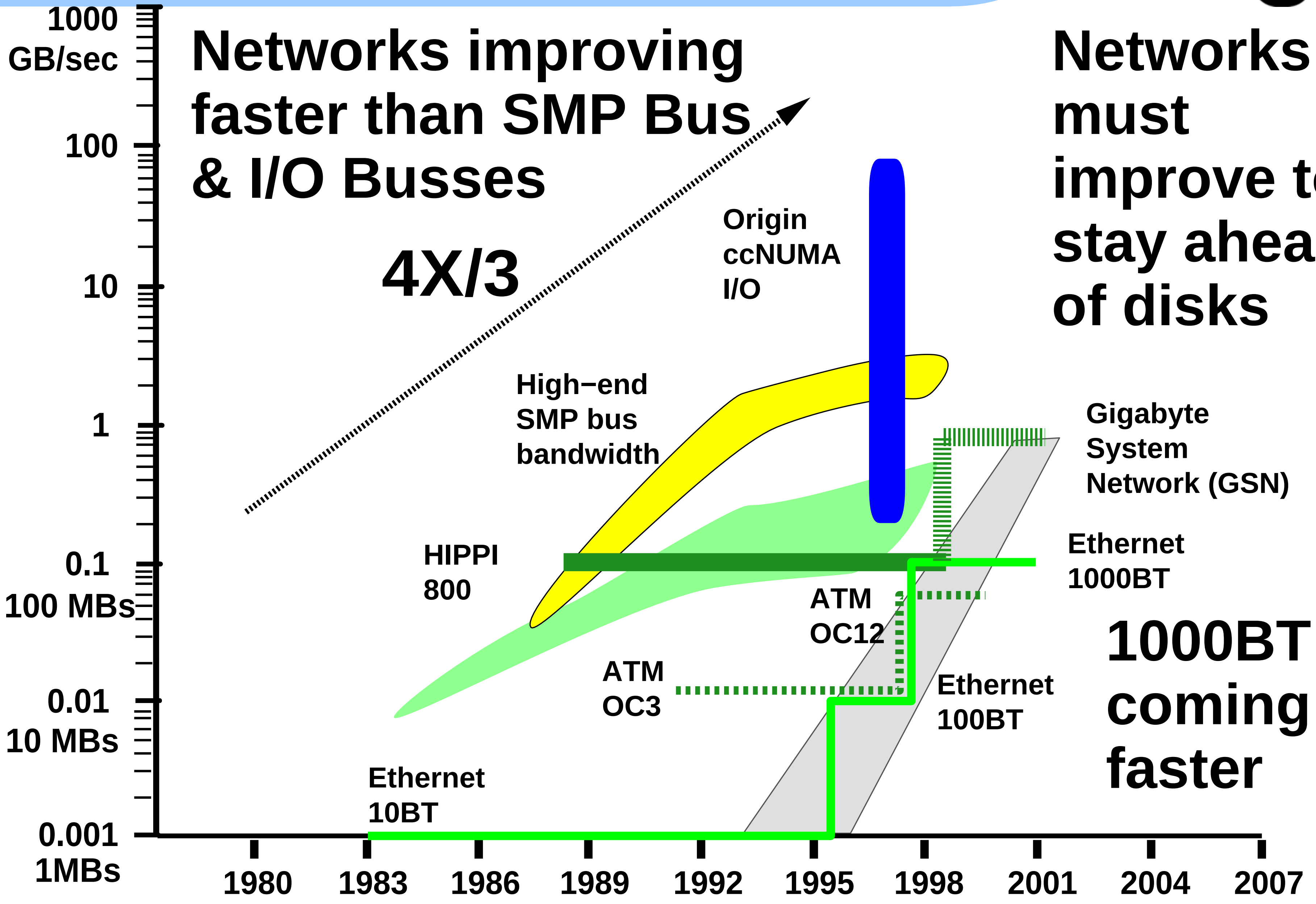
128p Origin, Onyx2:
up to
80GB/s I/O
40 GB/s memory,
20 GB/s Bisection

1 XIO, 1.28 GB/s

Origin200
PCI64,
.2 GB/s

Start small
Buy incrementally
Scale big

LAN, Interconnect Bandwidths

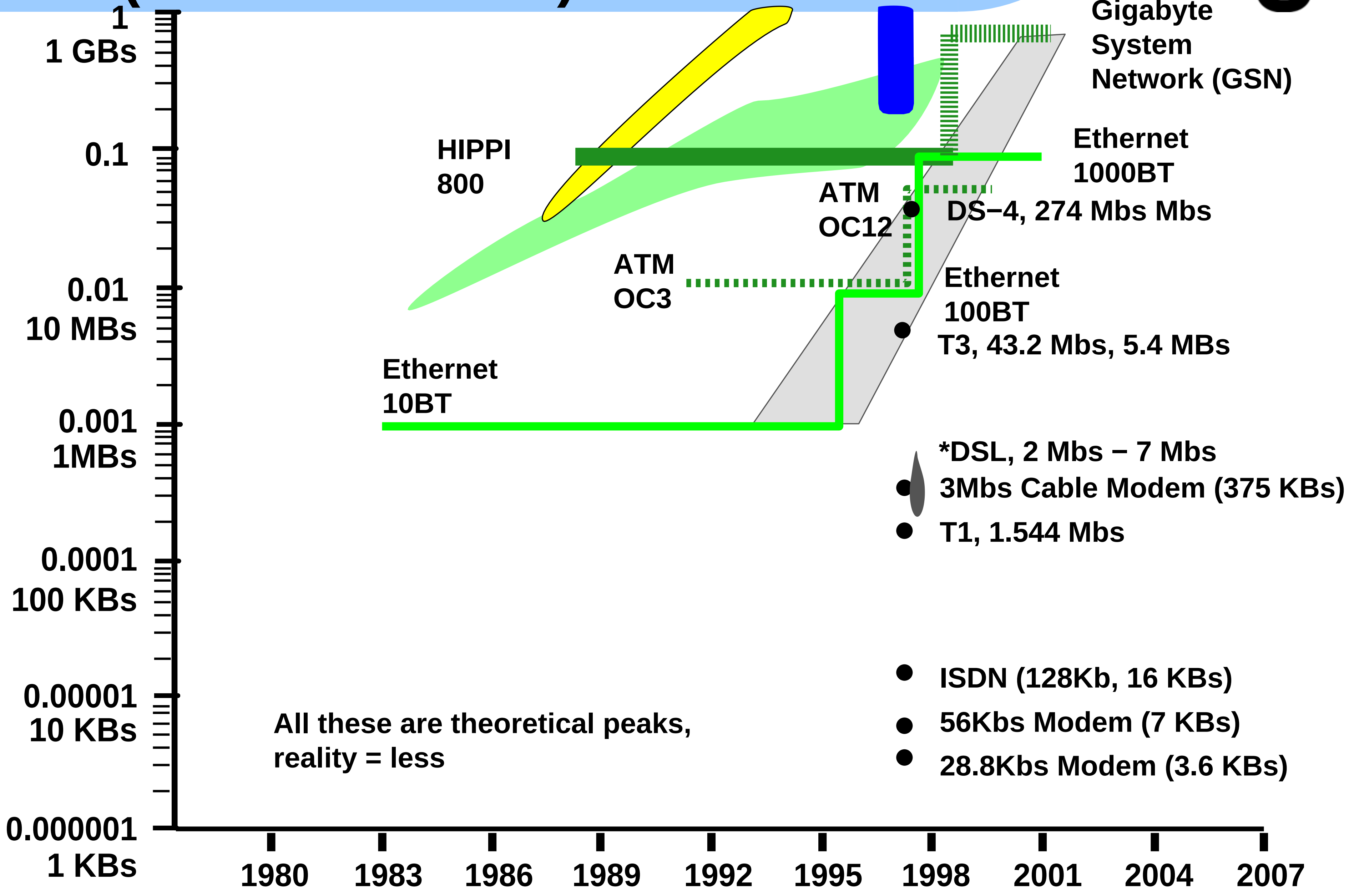


Networks improving faster than SMP Bus & I/O Busses

Networks must improve to stay ahead of disks

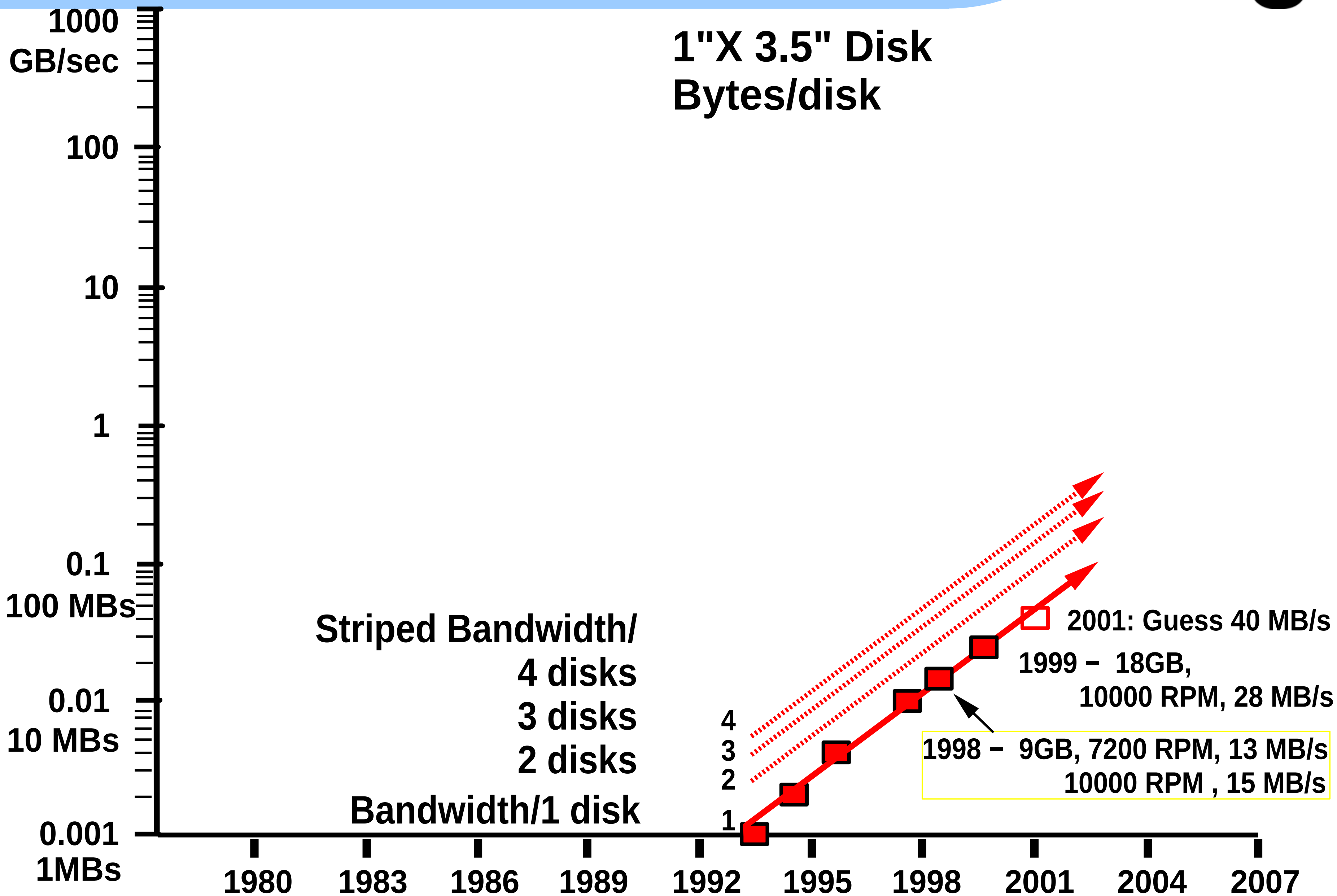
1000BT coming faster

Beyond the LAN
(Different Scale!)



All these are theoretical peaks,
reality = less

Disk Bandwidths (Highest)

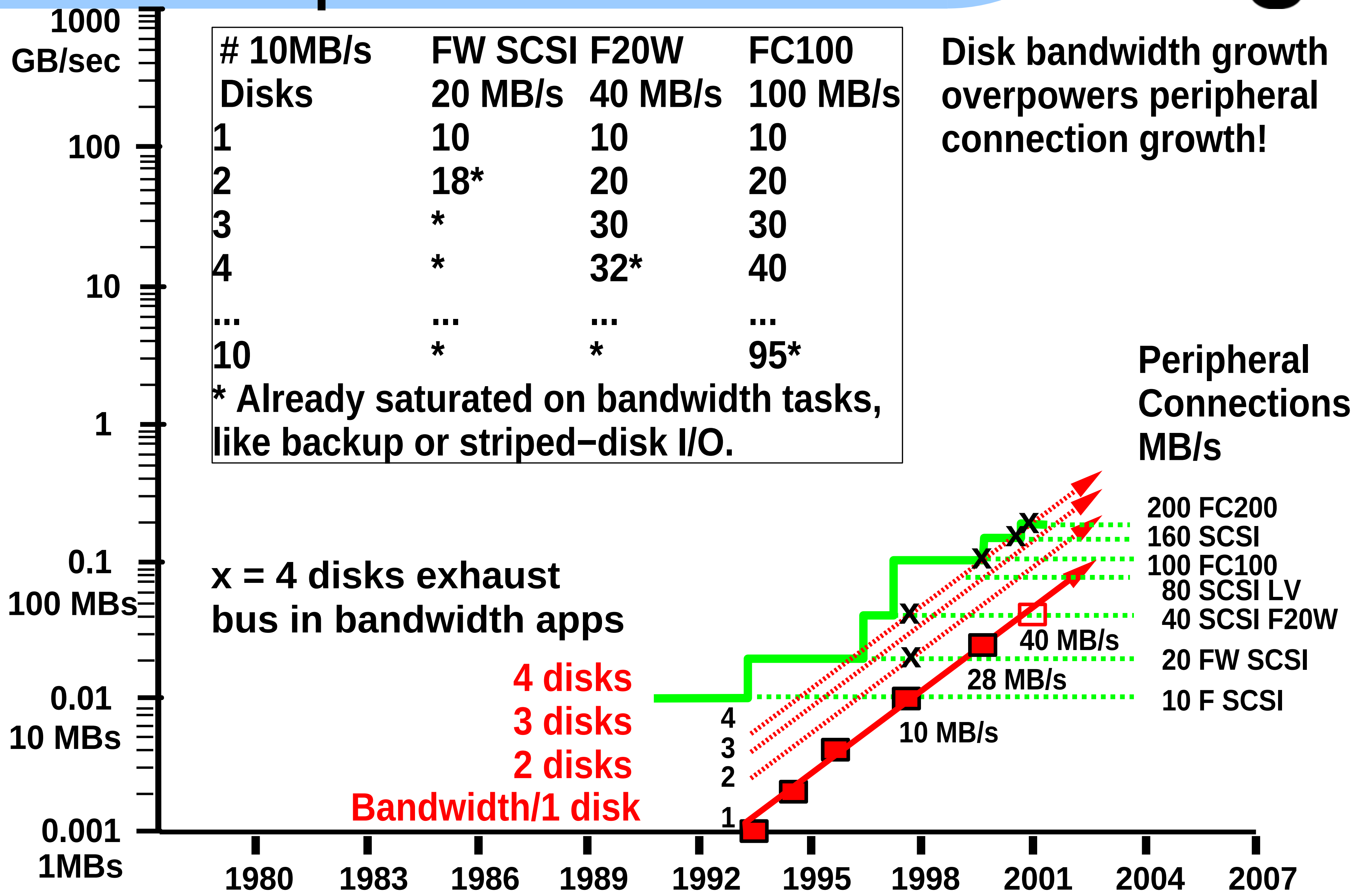


Fast Disk Bandwidth
vs Peripheral Connections

# 10MB/s Disks	FW SCSI 20 MB/s	F20W 40 MB/s	FC100 100 MB/s
1	10	10	10
2	18*	20	20
3	*	30	30
4	*	32*	40
...	...	...	...
10	*	*	95*

* Already saturated on bandwidth tasks, like backup or striped−disk I/O.

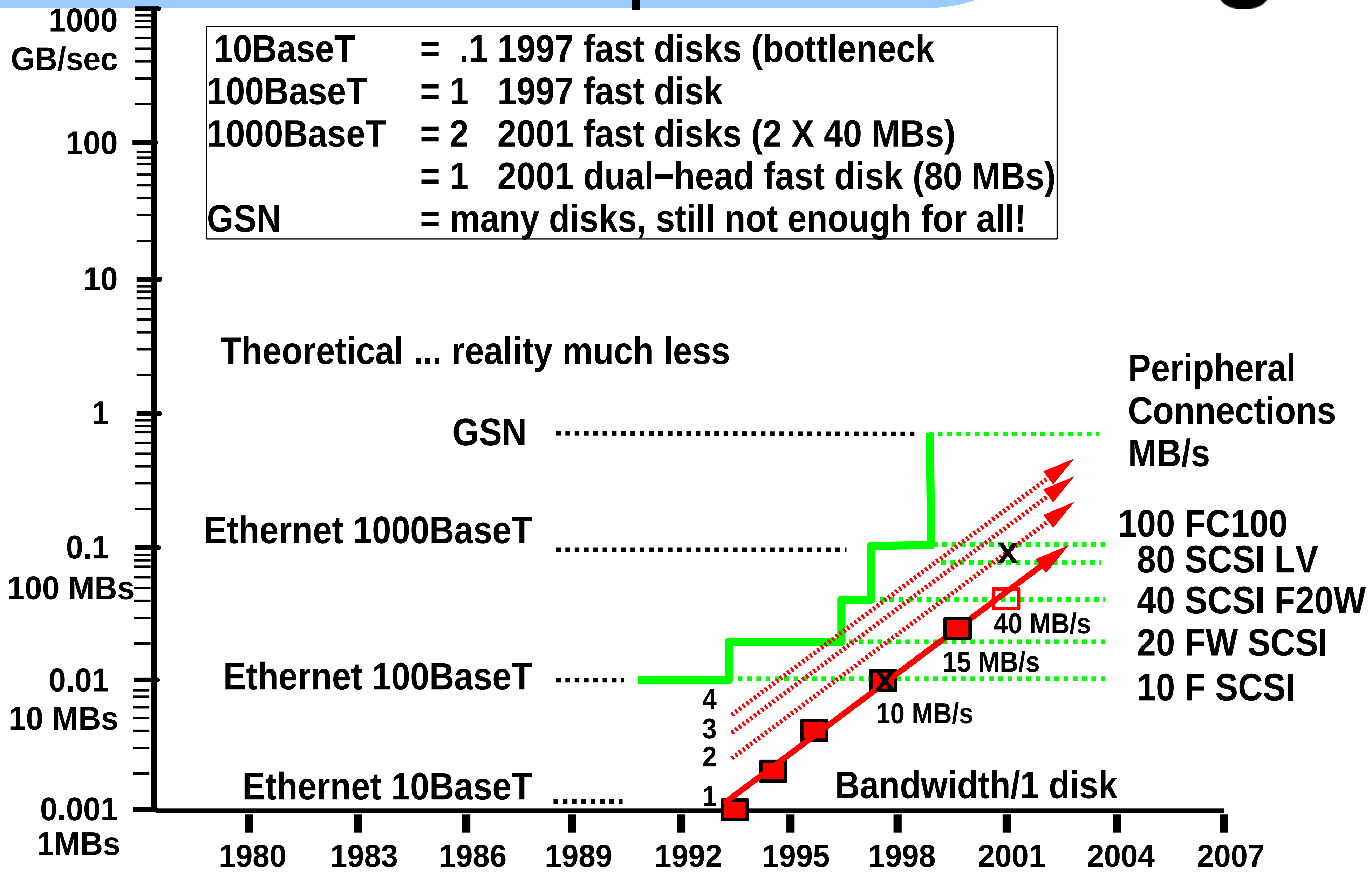
Disk bandwidth growth overpowers peripheral connection growth!



x = 4 disks exhaust bus in bandwidth apps

Fast Disk Bandwidth

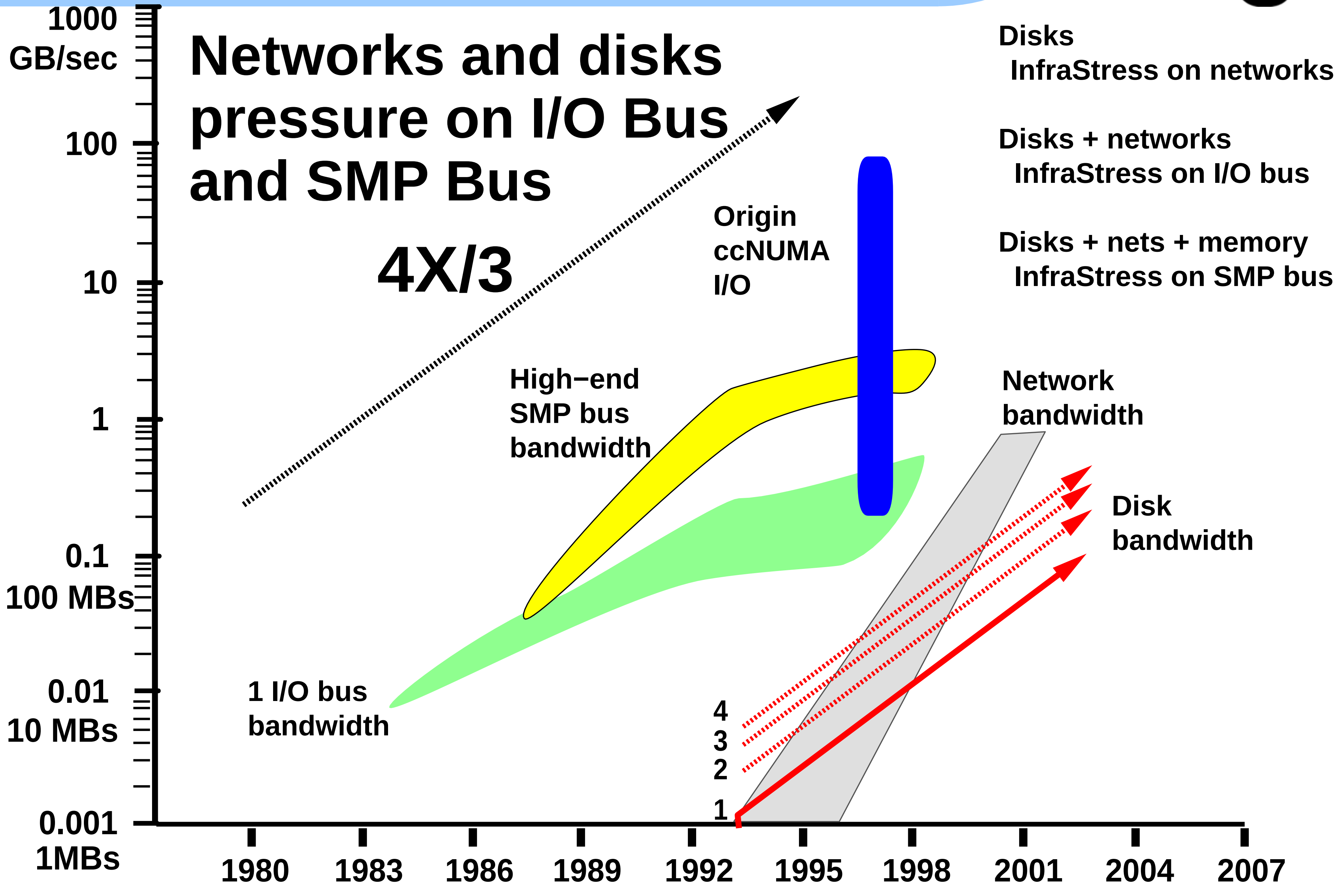
vs Networks & Peripheral Connections



10BaseT = .1 1997 fast disks (bottleneck
100BaseT = 1 1997 fast disk
1000BaseT = 2 2001 fast disks (2 X 40 MBs)
= 1 2001 dual−head fast disk (80 MBs)
GSN = many disks, still not enough for all!

Theoretical ... reality much less

Bandwidths – Summary



Networks and disks pressure on I/O Bus and SMP Bus

4X/3

Disks
InfraStress on networks

Disks + networks
InfraStress on I/O bus

Disks + nets + memory
InfraStress on SMP bus

Bandwidths – Implications

1. SMP Busses not growing with 4X/3
   Interconnect and memory bandwidth limits
   ==> Crossbars
     Centralized (mainframe)
     Distributed (ccNUMA)

2. Some I/O busses, peripheral connects,
   and especially networks under pressure
   to keep up with disk bandwidth

3. Disks are faster than tapes ... backup?

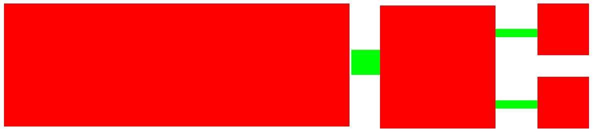
4. SANs for bandwidth and latency

Interactions: Distributed Data

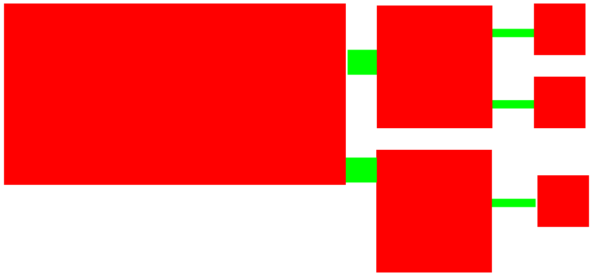
Shape of solution driven by shape of hardware?
"Natural" distribution of work: cost−effective
"Unnatural" data distribution: very painful
High bandwidth, low latency, or else...

Better: make hardware match shape of problem

Problem Shape

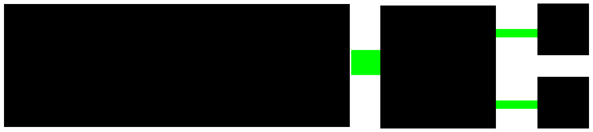


growth??

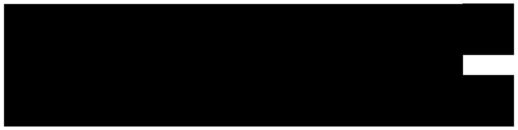


Solution Shape?

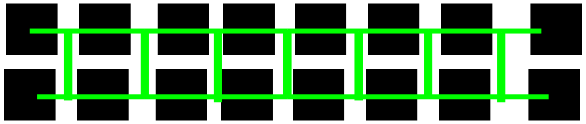
Good Fit
(technology)



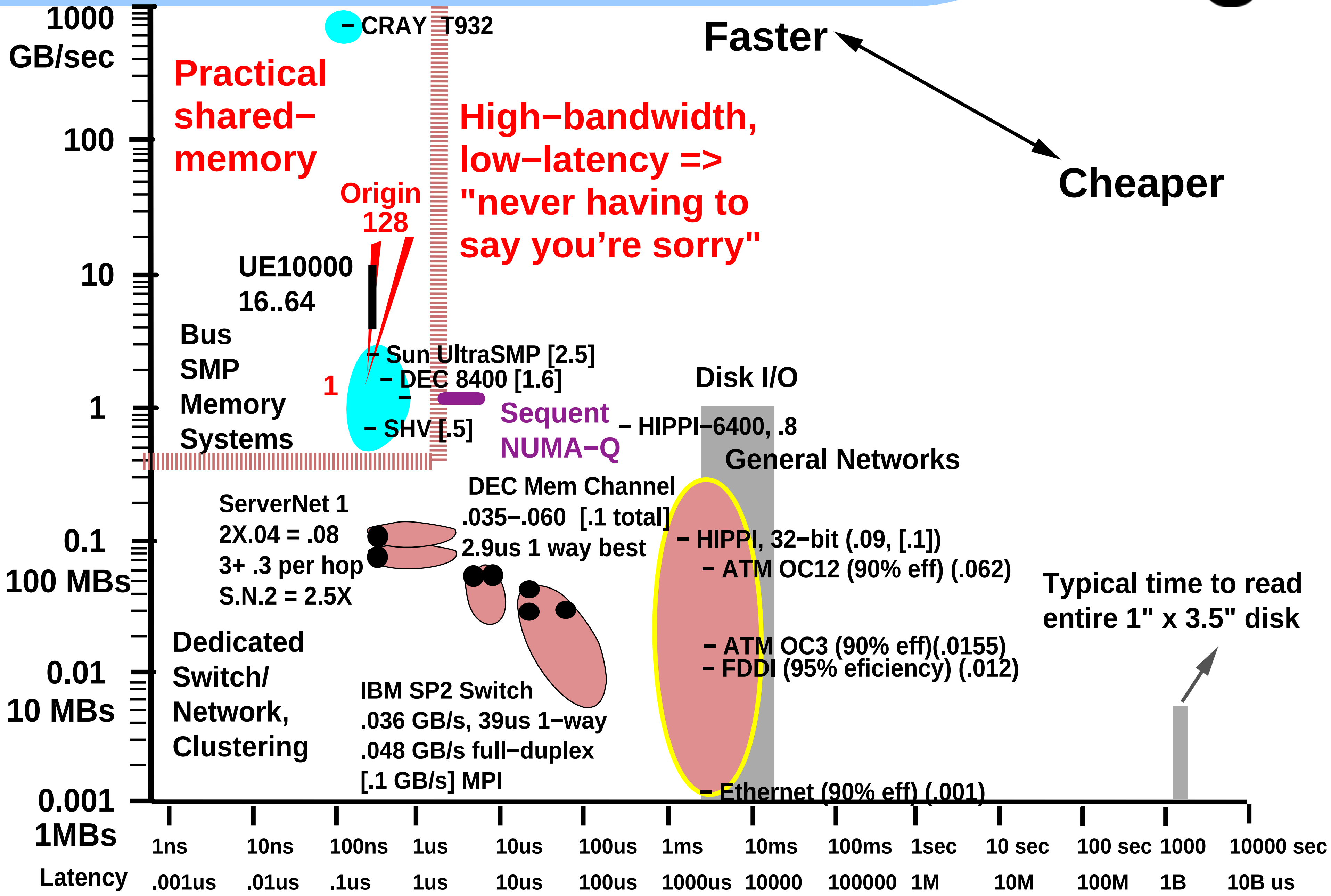
Centralize
(allocation)



Decentralize
(partitioning,
administration)



Interactions:
Bandwidths vs Latencies



Interactions:
Disk Technology Trends

- Capacities
  - Grow very fast
- Latencies
  - Barely improve for small blocks
  - Improve moderately for large blocks
- Bandwidths
  - Improve, but not so fast as capacity
  - Capacity/bandwidth ratios get worse
  - Pressure −> more smaller disks
- Interactions
  - 100BaseT, PCI32, F+W SCSI overrun
  - Backup rethinking
    - Desktop & 2 half−empty disks?
    - Backup servers?

Technology Summary

	Good	Bad	Ugly
CPU	Mhz	Parallelism	Latency
SRAM	On−chip		Latency
RAM	Capacity		Latency
Disk	Capacity		Latency
Tape	Capacity	Bandwidth	Latency
Network	Bandwidth		Latency
Software			Work!
Sysadmin	Technology		Exciting

Conclusion: InfraStress

Wishlist for Overcoming It

1. Find/understand: insight
   Tools: Navigate, organize, visualize
2. Input: creativity
   Tools: create content from ideas

3. Store and process the data: power
   Big addressing, modern file system
   Big I/O (number and individual speed)
   Big compute (HPC or commercial)
4. Move it: unleash
   Scalable interconnect
   High−performance networking

5. Change: survive!
   Incremental scalability, headroom
   Infrastructure already upgraded

References

1.http://www.storage.ibm.com/hardsoft/diskdrdl/library/technolo.htm
IBM storage web page