

# Data services – from data to containers

FAST 2003 keynote  
john wilkes

## Key messages



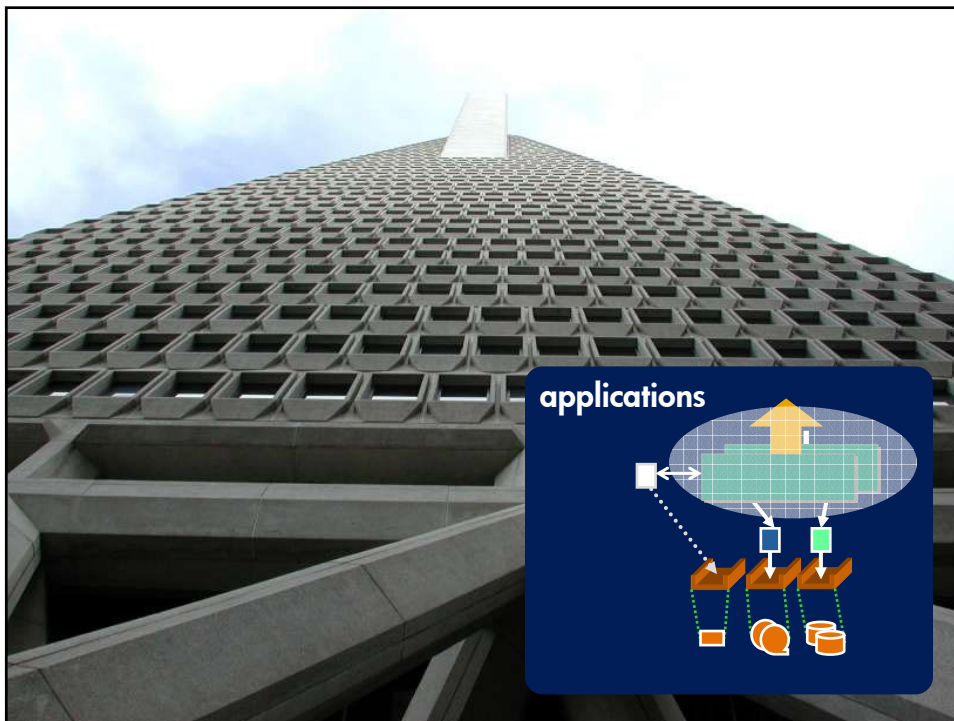
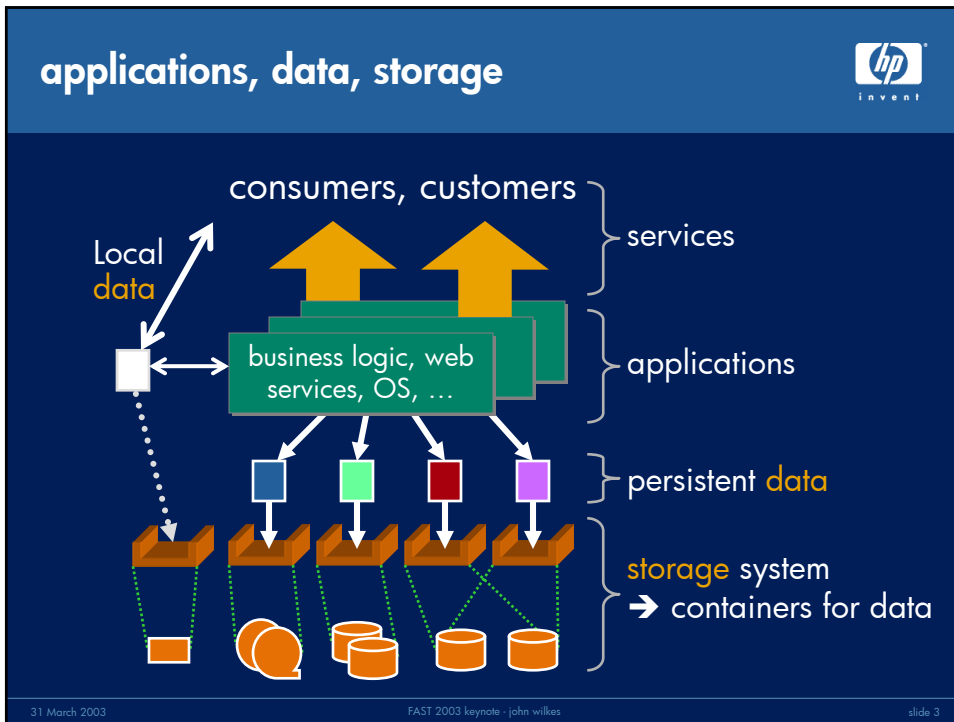
Rising **system complexity** +  
rising **abilities** +  
rising **expectations**

Solution:

- define **data QoS** needs
- use **storage QoS** abilities
- **automate** storage + data management

Our target should be:  
**data services**

31 March 2003 FAST 2003 keynote - john wilkes slide 2



## Personal applications



### The digital life

- Information on the move
- Data at home
- Interactions everywhere
- An information-hungry society

"When I get a little money, I buy books. And if there is any left over, I buy food."

– Erasmus



31 March 2003

FAST 2003 keynote - john wilkes

slide 5

## Personal applications



### Wherever I go ... there's my data

- from **islands of isolated data** (work, home, on the move, PC, laptop, PDA, server, ...)
- to **anywhere, anytime access to data**



31 March 2003

FAST 2003 keynote - john wilkes

slide 6

## Personal applications



1TB in my pocket!

### Now what?

- security?
- privacy?
- resiliency?
- freshness of data?
- relevance?
- validity?



31 March 2003

FAST 2003 keynote - john wilkes

slide 7

## Personal applications + the back-end



there is no middle!




31 March 2003

FAST 2003 keynote - john wilkes

slide 8


## Enterprise (commercial) applications

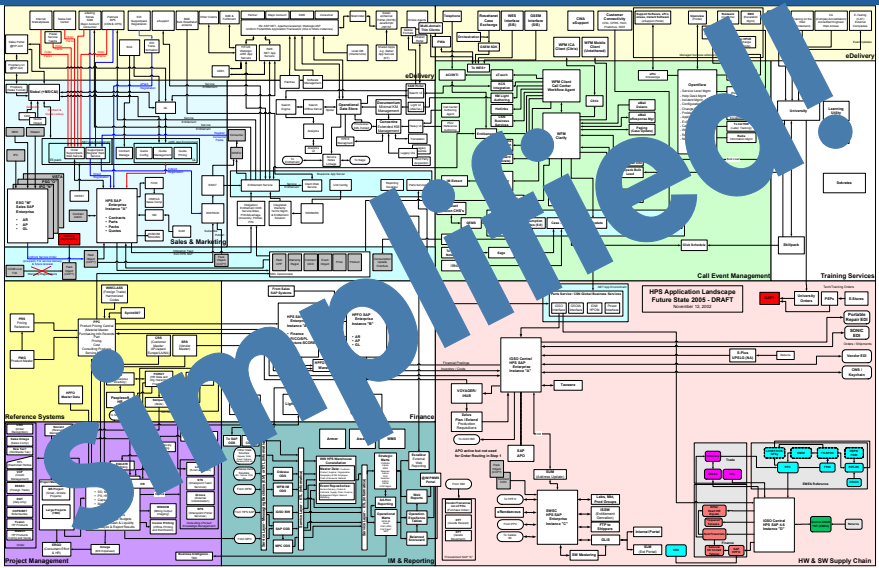


- on-line
  - “business critical” communications
    - email, workflow systems, ...
  - OLTP (e.g., order entry)
  - customer interactions (e.g., Verizon)
  
- back office
  - SAP, ERM, ...
  - day-to-day finance systems
  - logistics planning and operations
  - payroll, ...
  
- ...

31 March 2003
FAST 2003 keynote - john wilkes
slide 9

## sample enterprise IT plan





31 March 2003
FAST 2003 keynote - john wilkes
slide 10

these are not desktop systems!



31 March 2003

FAST 2003 keynote - john wilkes

slide 11

Enterprise – what's next?



**Scientific applications as predictors of future trends?**

- Huge quantities of data
  - instruments, sensors, time sequences, ...
- Data often:
  - stochastic (“merely representative”)
  - large-scale
  - specialized access patterns
- Some examples:  
bioinformatics, physics, enterprise data mining inputs

31 March 2003

FAST 2003 keynote - john wilkes

slide 12

# Scientific applications – Sanger



## Sequencing Facility



Information Technology - 2002



Graphics from: Phil Butcher, *Meeting user demands: a solution architecture*, <http://www.sanger.ac.uk/Info/IT/>, Sanger Institute, 2002

31 March 2003

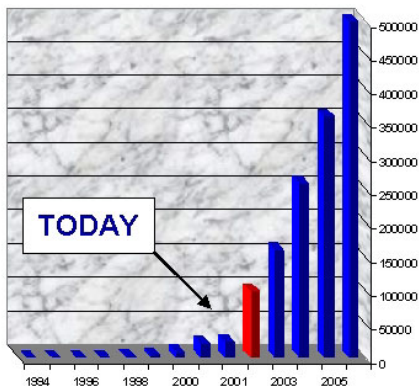
FAST 2003 keynote - john wilkes

slide 13

# Scientific applications – Sanger



## Data Storage



### FUTURE

To meet the scientific goals we believe we need to add around 80 - 100TB of storage each year for the next 4-5 years

Information Technology - 2002




Graphics from: Phil Butcher, *Meeting user demands: a solution architecture*, <http://www.sanger.ac.uk/Info/IT/>, Sanger Institute, 2002

31 March 2003

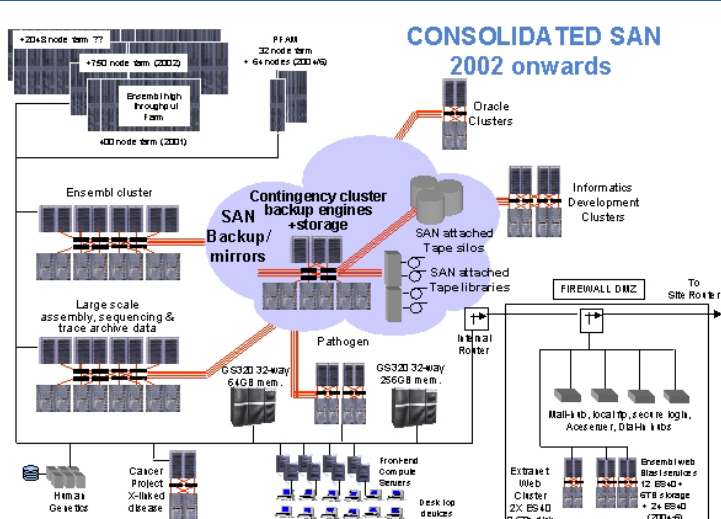
FAST 2003 keynote - john wilkes

slide 14

## Scientific applications – Sanger



### CONSOLIDATED SAN 2002 onwards




Graphics from: Phil Butcher, *Meeting user demands: a solution architecture*, <http://www.sanger.ac.uk/Info/IT/>, Sanger Institute, 2002

31 March 2003

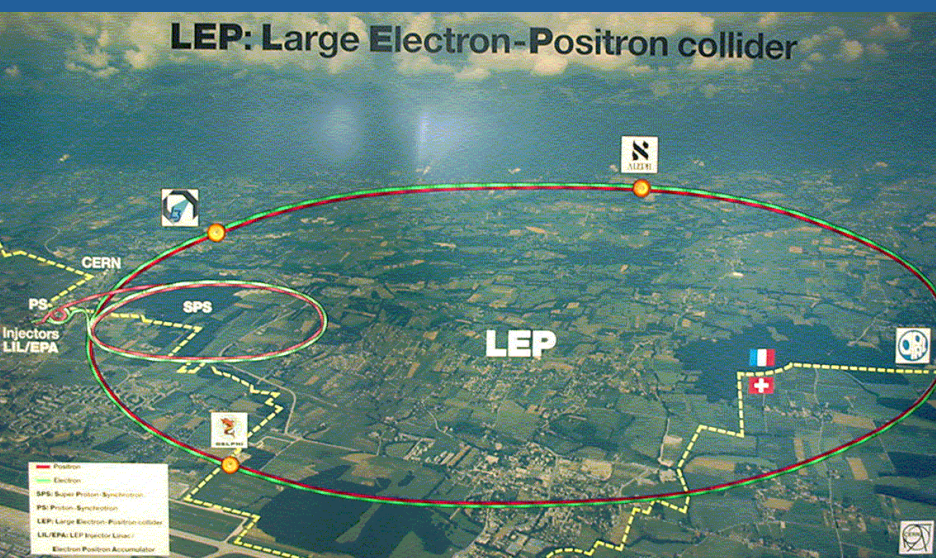
FAST 2003 keynote - john wilkes

slide 15

## Scientific applications – CERN



### LEP: Large Electron-Positron collider



31 March 2003

FAST 2003 keynote - john wilkes

slide 16

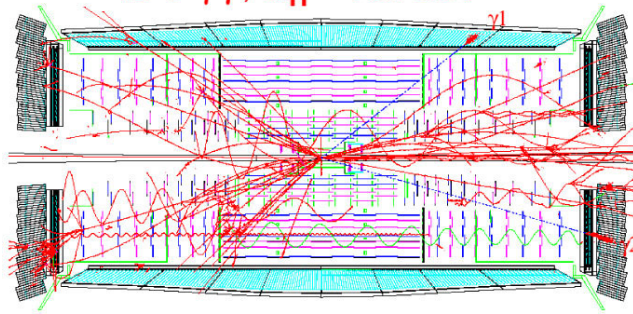


# Scientific applications – CERN



## Higgs event into two Photons

$H \rightarrow \gamma\gamma, M_H = 100 \text{ GeV}$



CMS February 2001

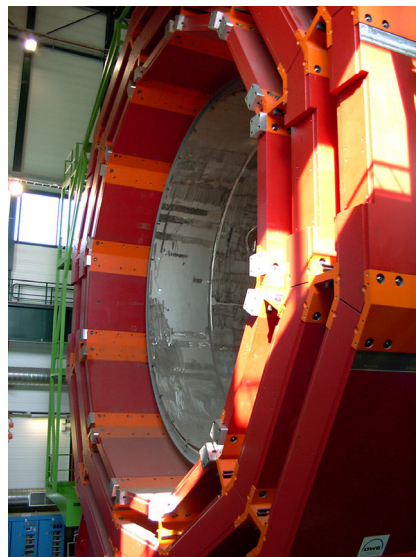
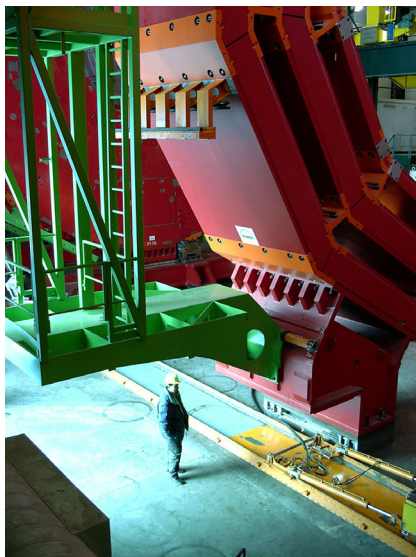
21

31 March 2003

FAST 2003 keynote - john wilkes

slide 17

# Scientific applications – CERN




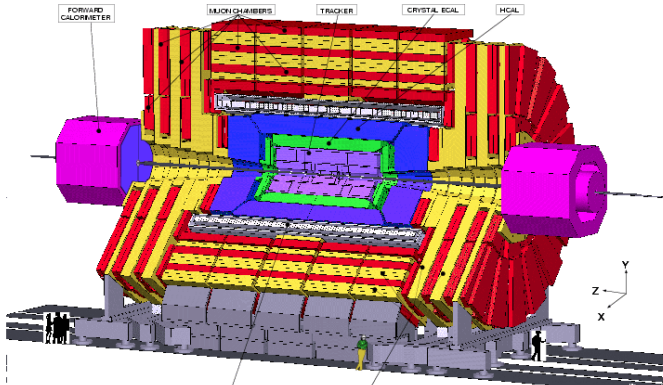
31 March 2003

FAST 2003 keynote - john wilkes

slide 18

## Scientific applications – CERN






Total Weight : 12,500t  
 Overall Diameter : 15.00m  
 Overall Length : 21.60m  
 Magnetic Field : 3T Tesla

FORWARD CALORIMETER  
 MUON CHAMBERS  
 TRACKER  
 CRYSTAL ECAL  
 HCAL  
 SUPERCONDUCTING COIL  
 RETURN YOKE

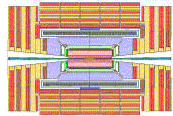
CMB-PARA-001-200697 PP

31 March 2003
FAST 2003 keynote - john wilkes
slide 19

## Scientific applications – CERN

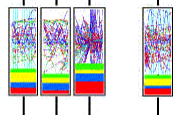


detector proper  
40MHz collisions




**1MB/event**

50,000 data channels  
200 GB buffering

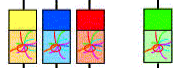


**~1TB/s**




**~500Gb/s**

event filtering  
(1 CPU/event)



**~0.5GB/s**

**data storage**



**~5PB/year**

31 March 2003
FAST 2003 keynote - john wilkes
slide 20


# Scientific applications – data grids

Today, CERN has 4000–6000 active clients, 2000 of which are offsite. Tomorrow, they will have ~ten tier1 data/computation centers spread across the globe.

31 March 2003 FAST 2003 keynote - john wilkes slide 21

data


## Data versus storage (containers)



data my preferences  
 container user keystroke history log  
 ↓ *application*  
 data user keystroke history  
 container file (byte vector)  
 ↓ *file system*  
 data file (named)  
 container volume (virtual block vector)  
 ↓ *volume virtualization system*  
 data volume fragment  
 container LU  
 ↓ *disk array*  
 data LU fragment  
 container disk drive

31 March 2003 FAST 2003 keynote - john wilkes slide 23

## Data attributes



- what's **true** about the data?
  - how much?
  - rate of growth?
  - access patterns?
- what **do we want to be true** about the data?
  - as above, plus ...
  - resiliency?
  - security?
  - semantics?
- plus ... **predictability** in the face of change

31 March 2003 FAST 2003 keynote - john wilkes slide 24

## Data attributes



### All data not created equal

Some data:

- has little value
- never changes
- can be regenerated



Not all data needs to be:

- accessible
- kept forever
- secret
- up to date



31 March 2003

FAST 2003 keynote - john wilkes

slide 25

## Data attributes – QoS



- **size**
- **access**
  - from where?
  - how fast?
  - when? (expectations; remote vs wired)
- **resilience**
  - data loss/corruption  
(operator error, software bugs, viruses, ...)
- **security**
  - who can access/change/control?
- **semantics**
  - consistency? updates? correctness?

**SLO:** a set of QoS objectives

**SLA:** a contract to provide them (adds penalties, monitoring rules, etc)

31 March 2003

FAST 2003 keynote - john wilkes

slide 26

## Data attributes - size



### the data itself

- current size
- over time
  - growth rates?
  - variance?
- unwanted data?



31 March 2003

FAST 2003 keynote - john wilkes

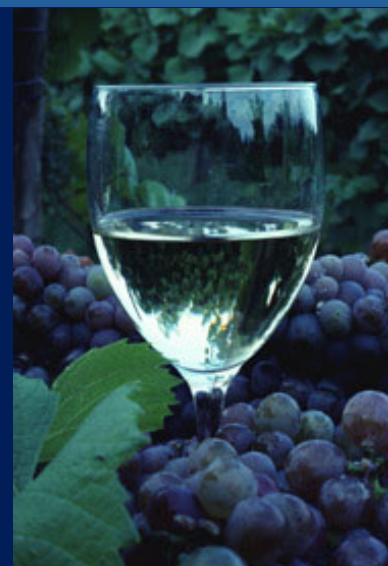
slide 27

## Data attributes - size



### after it's packaged

- wastage?
- slack capacity?
- duplicates
- duplicates




31 March 2003


FAST 2003 keynote - john wilkes

slide 28

## Data attributes – access




- from where?
  - local SAN/LAN
  - wired WAN
  - disconnected/mobile
- how fast?
  - QoS parameters
- when?
  - “availability”

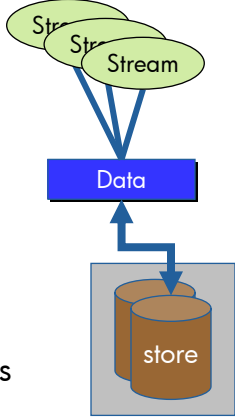


31 March 2003 FAST 2003 keynote - john wilkes slide 29

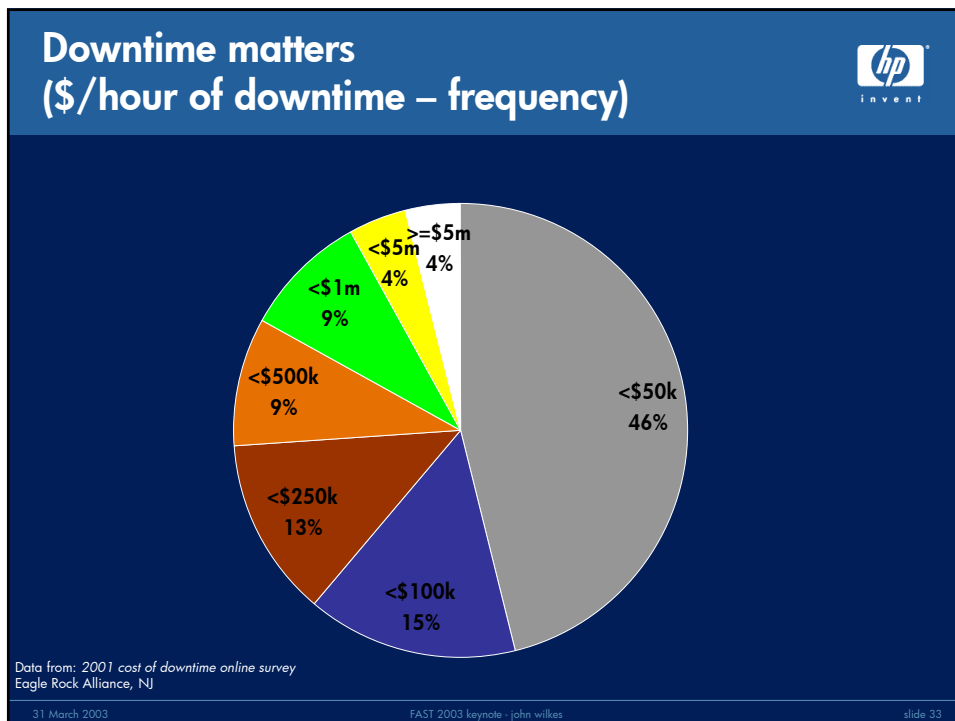
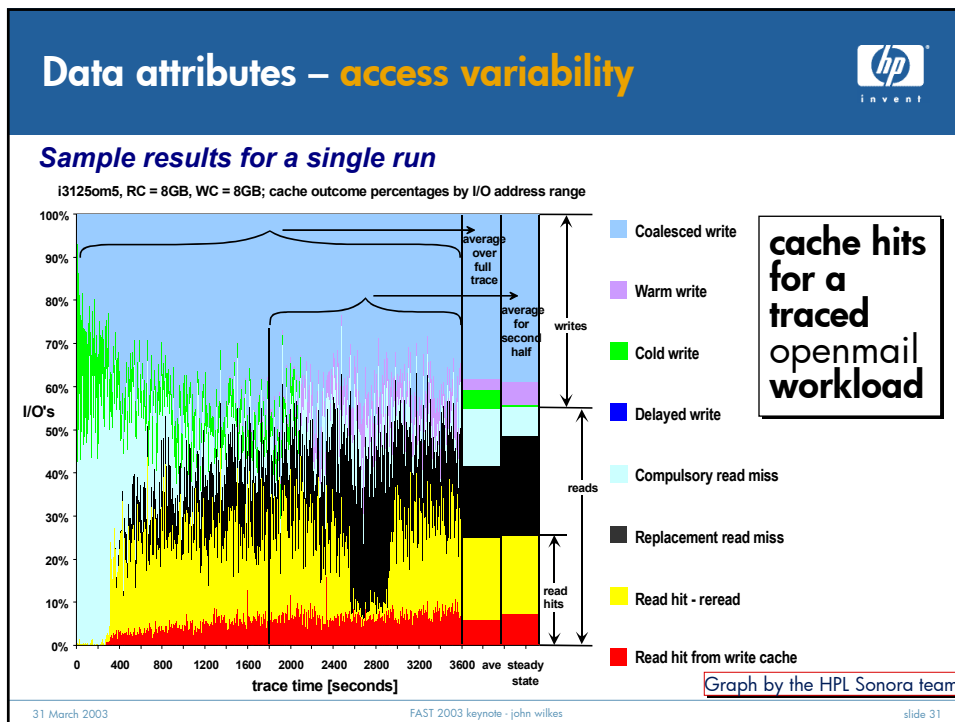
## Data attributes – access



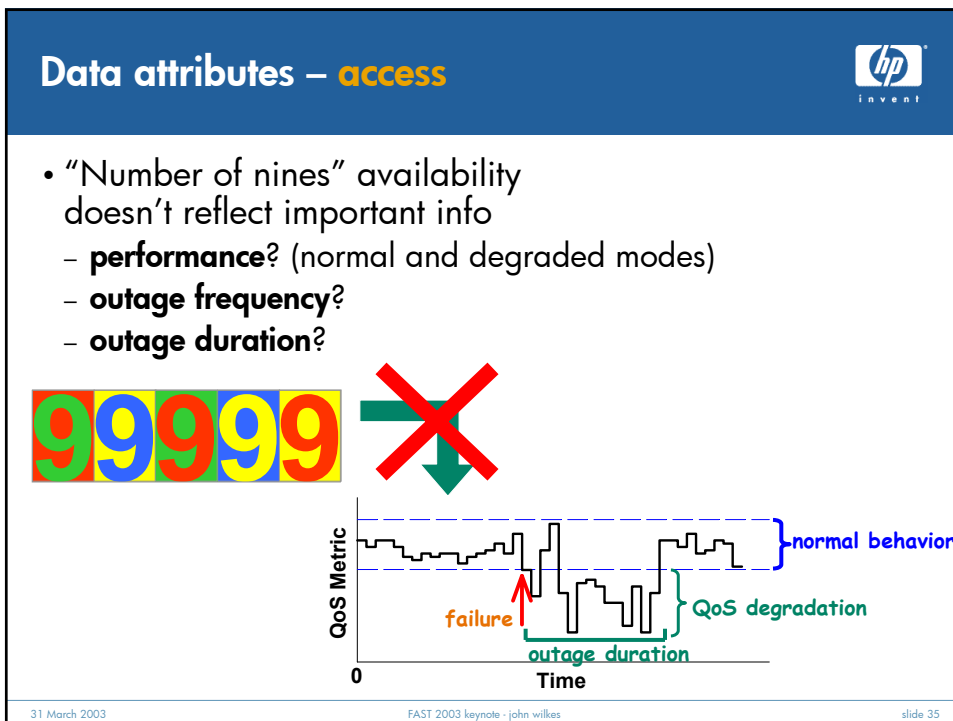
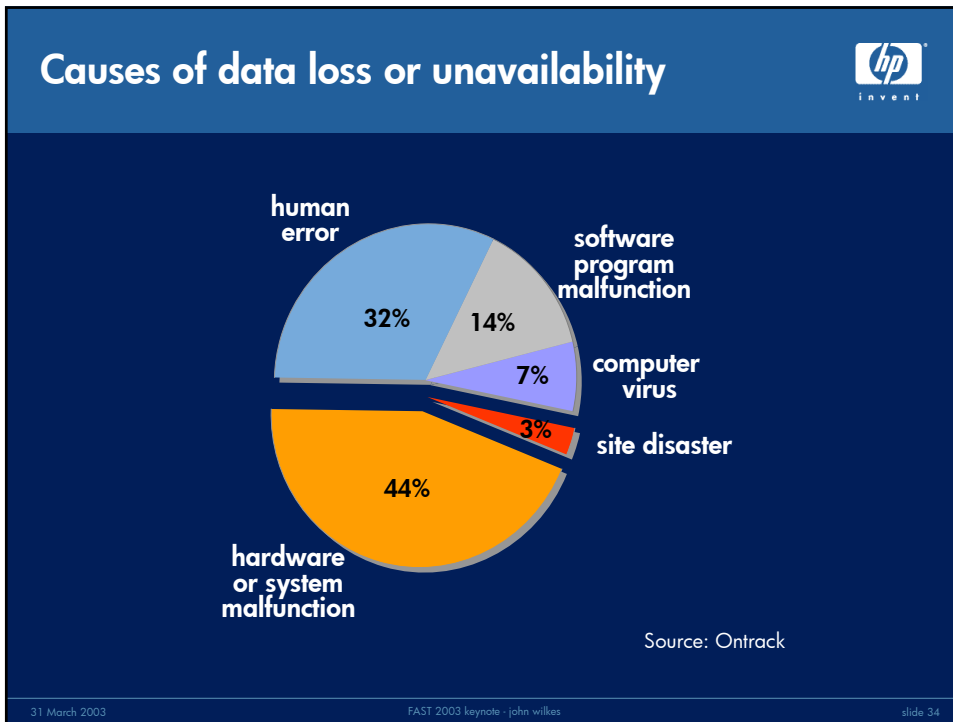
- **stream:** an access pattern
  - **behaviors:** I/O rate, I/O request size, spatial locality, temporal locality, cache affinity, phasing, ...
  - **requirements:** bandwidth, response time
- **data:** information being accessed
- **store:** a container
  - e.g., Logical Unit, Logical Volume
  - provider of resources to realize demands (capacity, bandwidth, ...)



31 March 2003 FAST 2003 keynote - john wilkes slide 30



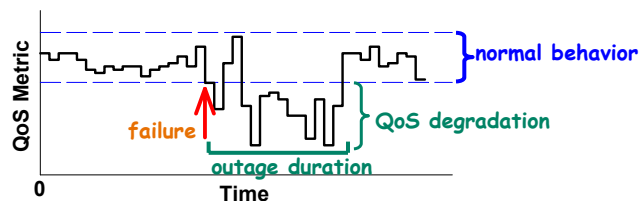




## Data attributes – access



- **Reliability:** likelihood system **up continuously** from 0 to t
- **Availability:** likelihood system **will be up** at time t
- **Performability:** likelihood system **has performance p** at time t



31 March 2003

FAST 2003 keynote - john wilkes

slide 36

## Data attributes – resilience/reliability



### Resilience/reliability

- **lack of data loss or corruption**, from:
  - operator error, software bugs, viruses, ...
  - hardware failures (a container property)
- simple model: **annual failure rate**
  - $AFR = 1/MTTDL$  ("mean time to data loss")
  - any size loss is equally bad
- a richer model:
  - < size [bytes], type [recent/arbitrary], rate [per year]>

31 March 2003

FAST 2003 keynote - john wilkes

slide 38

## Data attributes – resilience/reliability



### Recent data loss

- **usage** is driven by:
  - efficiency needs (e.g., buffering)
  - user's conceptual model ("yesterday's state")
- **recovery**
  - return to/retrieve data from a time-based **recovery point**
  - **re-apply** appropriate intermediate changes

31 March 2003

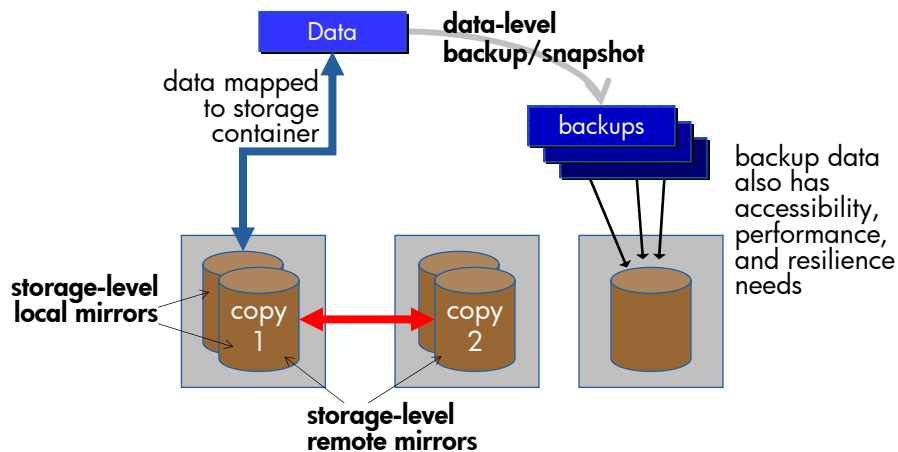
FAST 2003 keynote - john wilkes

slide 39

## Data attributes – resilience/reliability



### Implementation techniques



31 March 2003

FAST 2003 keynote - john wilkes

slide 40

## Data attributes – security



### Who can access/change/control data?

- implementation:
  - physical security (common today in FibreChannel SANs)
  - host, network, **storage device**
- secure **transmission**
- secure **storage**

31 March 2003

FAST 2003 keynote - john wilkes

slide 41

## Data attributes – semantics




- correctness
  - an information-level construct: not knowable at the data level
- consistency/coherency between versions
  - “two-phase locking reduces inconsistencies to a dull roar”*
  - **data versions** (e.g., backups, snapshots, archive)
  - **data copies** (e.g., software release)
  - **storage copies** (e.g., local/remote mirrors)

31 March 2003

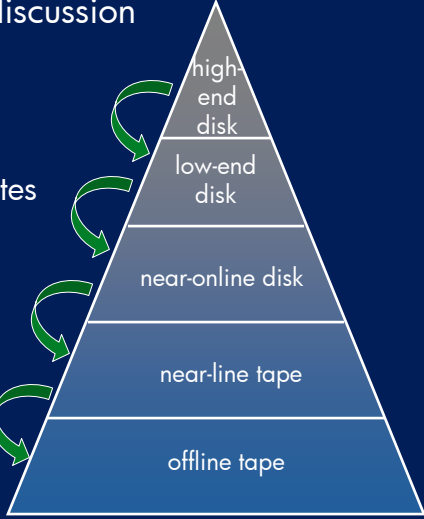
FAST 2003 keynote - john wilkes

slide 42

## Data life cycle and data placement




- Overlaid on all the previous discussion
- Sample **phases**:
  - gathered, generated
  - production use – access/updates
  - demotion/archiving
  - discarding/expunging
- All phases can exploit storage (container) hierarchy



31 March 2003 FAST 2003 keynote - john wilkes slide 43



## Storage



- disk drives
- MEMS, MRAM
- disk arrays
  
- tape drives
- tape libraries
  
- storage systems (SANs)
- storage management systems


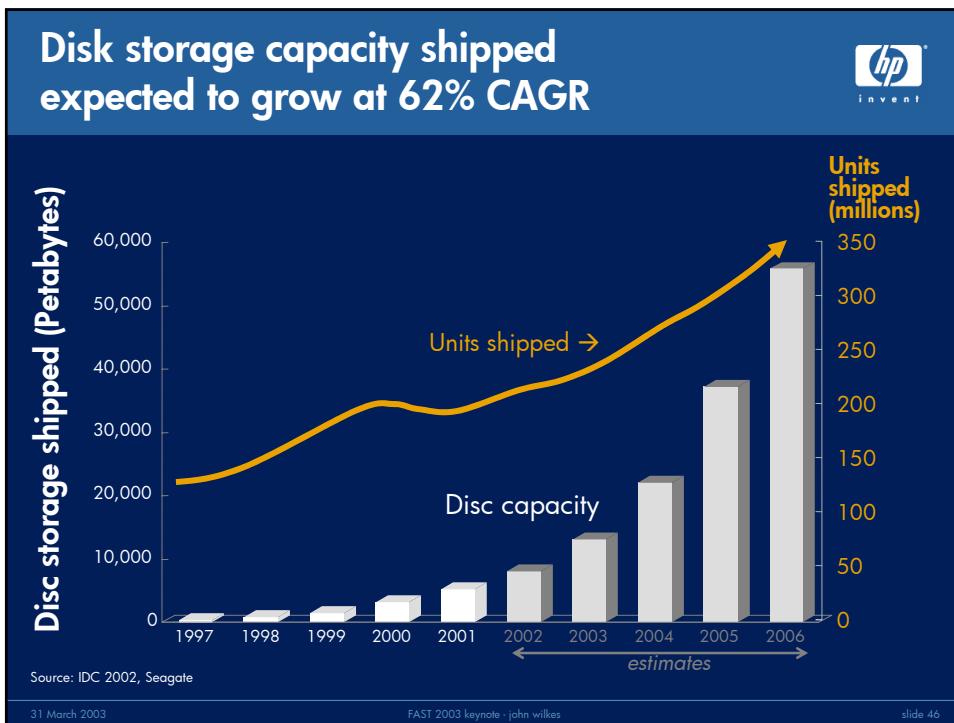
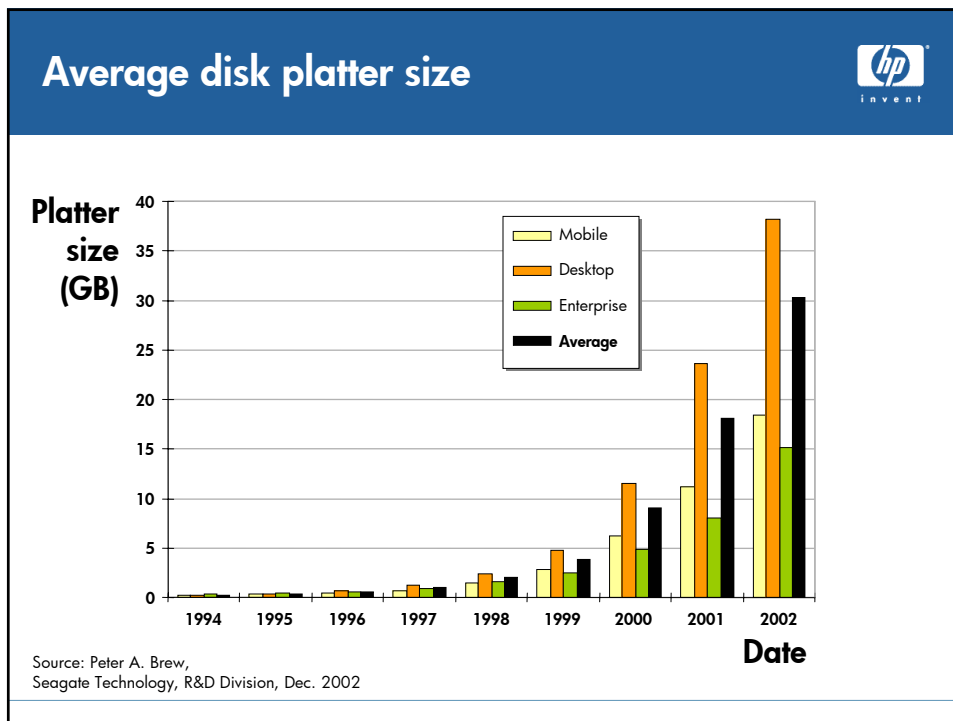
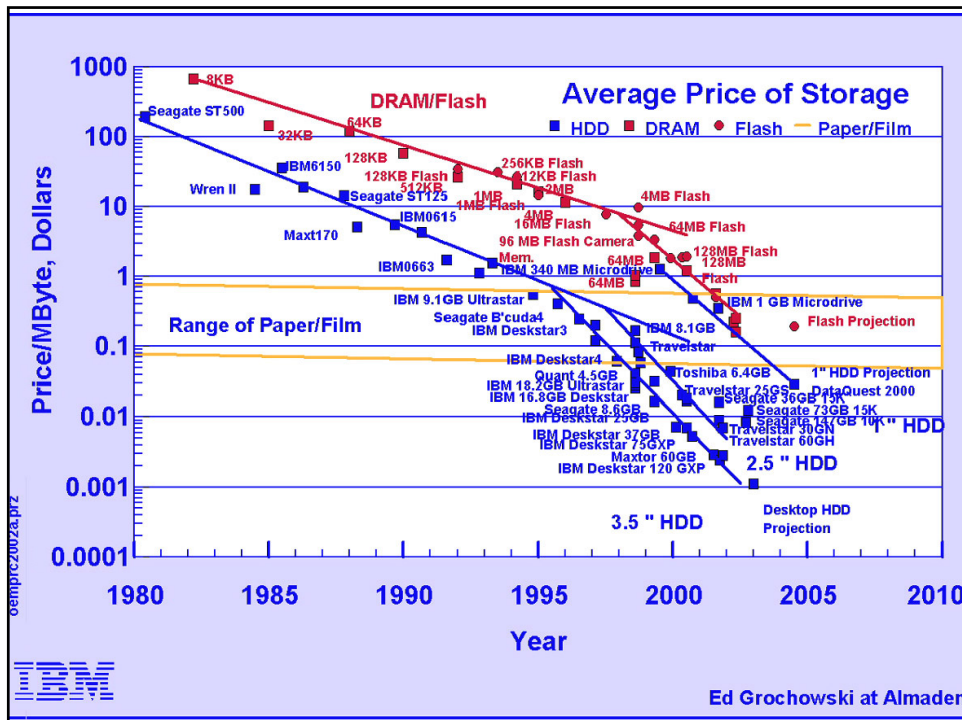
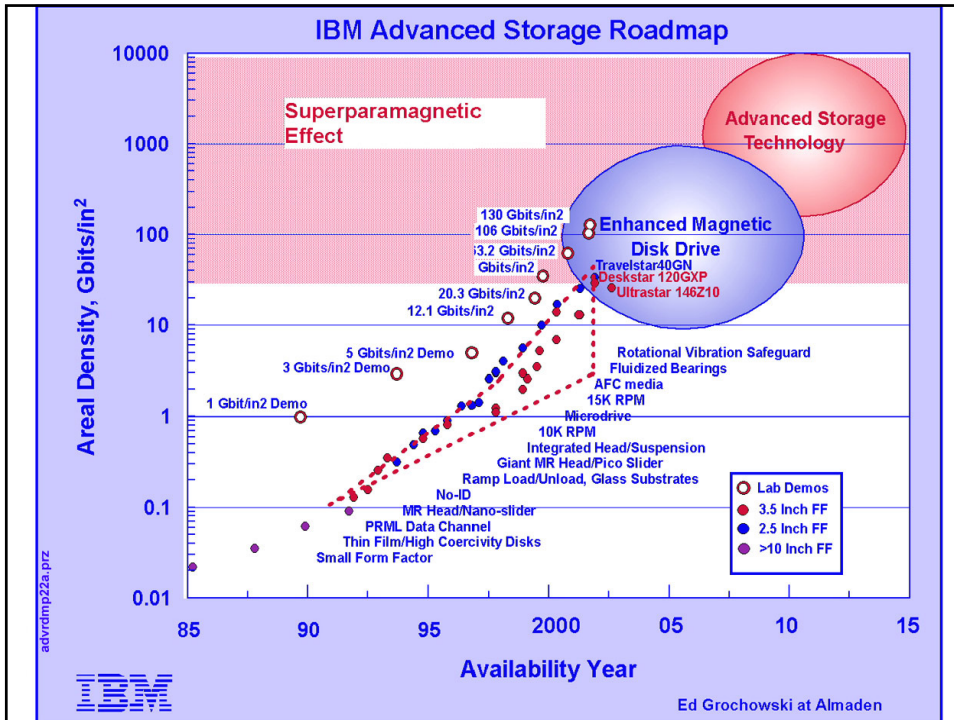


Image courtesy of Seagate Technology, Inc.  
© 2000 Seagate Technology, Inc.

31 March 2003
FAST 2003 keynote - john wilkes
slide 45







## Storage devices – MEMS

FIELD EMISSION TIP

Electron beam

EMISSION-TIP ARRAY

STORAGE MEDIUM

SUSPENSION SPRINGS

~ 30  $\mu\text{m}$  (about the diameter of a human hair)


~ 10x faster than a disk drive  
Cost < flash RAM  
Gbytes per module  
Nonvolatile  
Portable: small, rugged, low power

From: "Scientific American" Magazine  
May 2000 Issue, Page 72

31 March 2003 FAST 2003 keynote - john wilkes slide 50

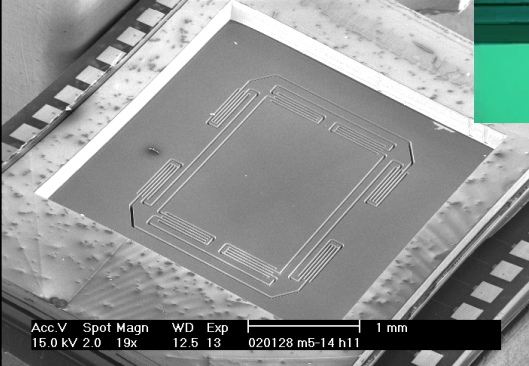


## Storage devices – MEMS

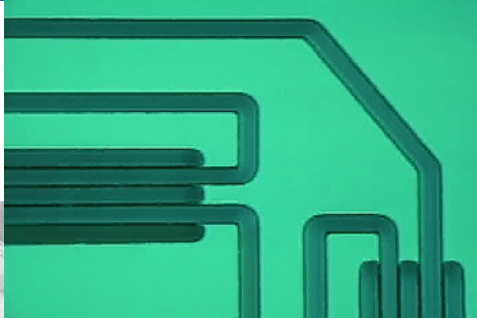


**Media-mover prototype**

- replaces spinning disks with a micro-machined motor
- faster access at lower power and lower cost




Acc.V 15.0 kV Spot Magn 2.0 19x W/D Exp 12.5 13 020128 m5-14 h11 1 mm



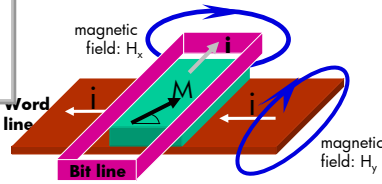
From: Jim Brug and Rich Elder, HP Labs

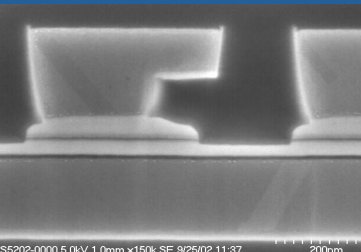
31 March 2003
FAST 2003 keynote - john wilkes
slide 51

## Magnetic RAM (MRAM) technology



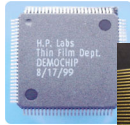
~ DRAM speeds  
< DRAM cost  
100's MB/chip  
non-volatile



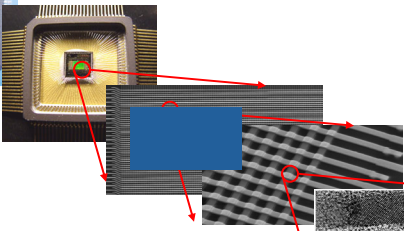


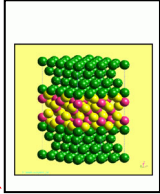
S5202-0000 5.0kV 1.0mm x150k SE 9/25/02 11:37 200nm

**Cross-section of memory cell**



H.P. Labs  
Thin Film Dept.  
DEMOCHIP  
8/11/99




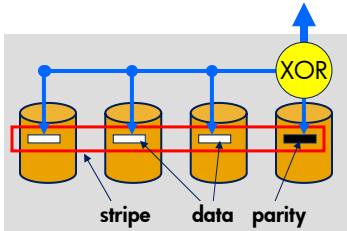


Images from: Jim Brug and Rich Elder, HP Labs


31 March 2003
FAST 2003 keynote - john wilkes
slide 52

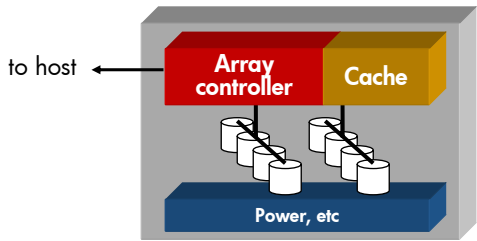
## Storage devices – disk arrays





stripe data parity XOR






to host

Array controller Cache

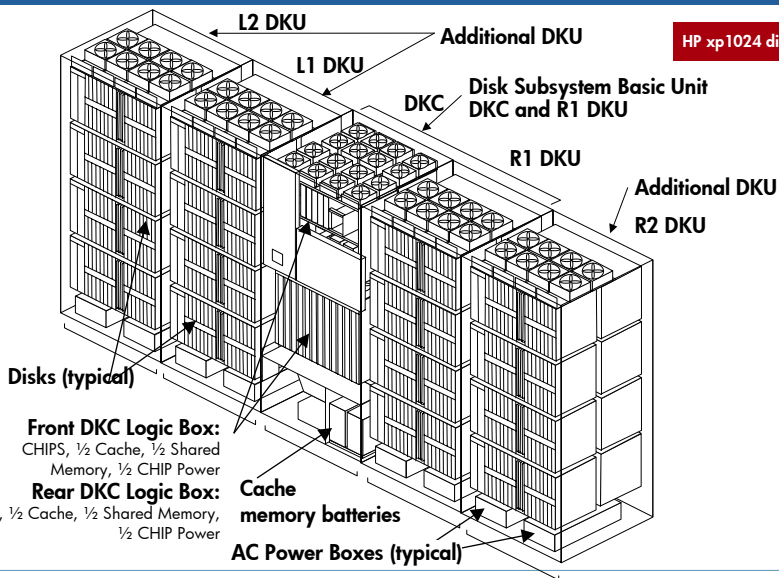
Power, etc

31 March 2003
FAST 2003 keynote - john wilkes
slide 53

## High-end disk array – physical structure



HP xp1024 disk array



L2 DKU Additional DKU

L1 DKU

Disk Subsystem Basic Unit DKC and R1 DKU

DKC

R1 DKU

Additional DKU R2 DKU

Disks (typical)

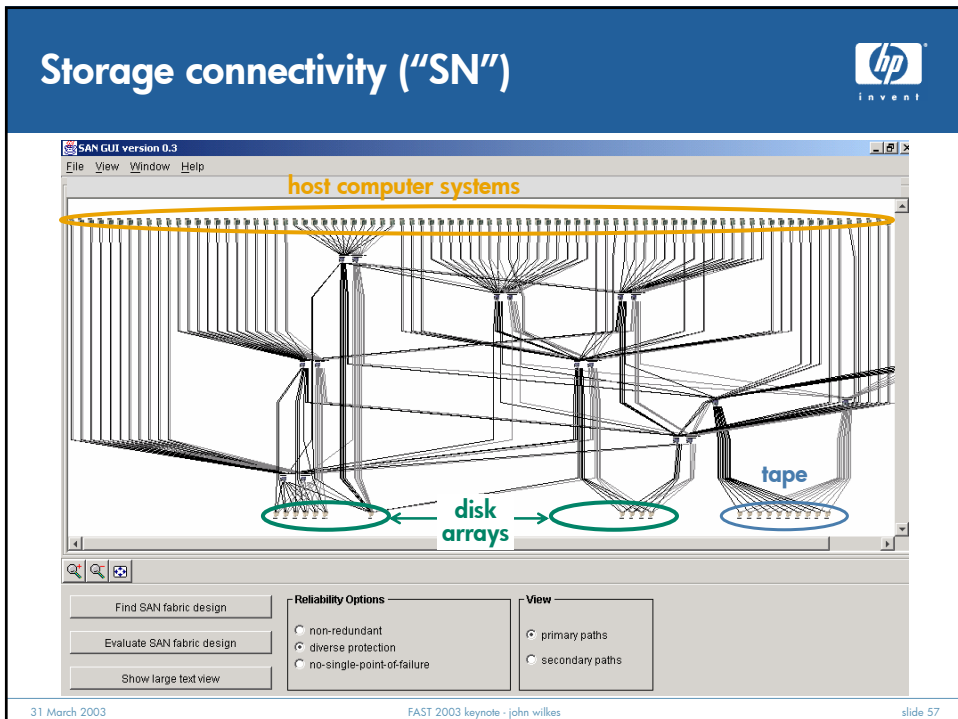
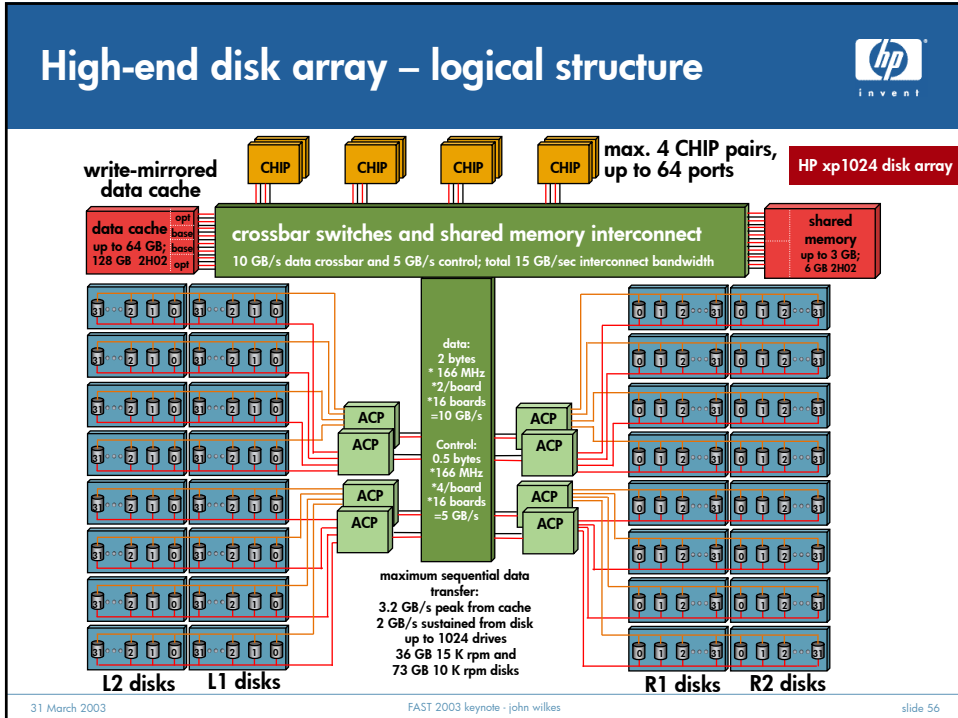
Front DKC Logic Box:  
CHIPS, 1/2 Cache, 1/2 Shared Memory, 1/2 CHIP Power

Rear DKC Logic Box:  
ACPs, 1/2 Cache, 1/2 Shared Memory, 1/2 CHIP Power

Cache memory batteries

AC Power Boxes (typical)

31 March 2003
FAST 2003 keynote - john wilkes
slide 55



## Disk systems getting smarter – and collective

**1988 john wilkes – DataMesh**

storage module: processor, Mechanism

LAN module: processor, LAN driver

**1999+ Jim Gray – “disks are becoming computers” (Storage bricks have arrived, FAST 2002)**

- Applications: Web, DBMS, Files, OS
- Disk controller + 1Ghz CPU + 1GB RAM
- Communications: Infiniband, Ethernet, radio...

**1996 Garth Gibson – NASD**

a) Current Trident ASIC 74 mm² at 0.68 micron

b) Next generation ASIC @ 35 micron technology

- features 100 Kbytes
- 7 controllers
- 7 Network support

35 micron traces 4.8 mm²

Insert: 35 micron @ on-gem RISC uP

fits in 27 mm² with BK + BK cache at 200 Mhz, 230 Dhrystone MIPS

**2001 IBM Almaden – Collective Intelligent bricks (aka IceCube)**

Brickwall, Cube, Rack-like

**2003 HPL – federated array of bricks (FAB)**

31 March 2003 FAST 2003 keynote - john wilkes slide 58

## The SNIA shared storage model

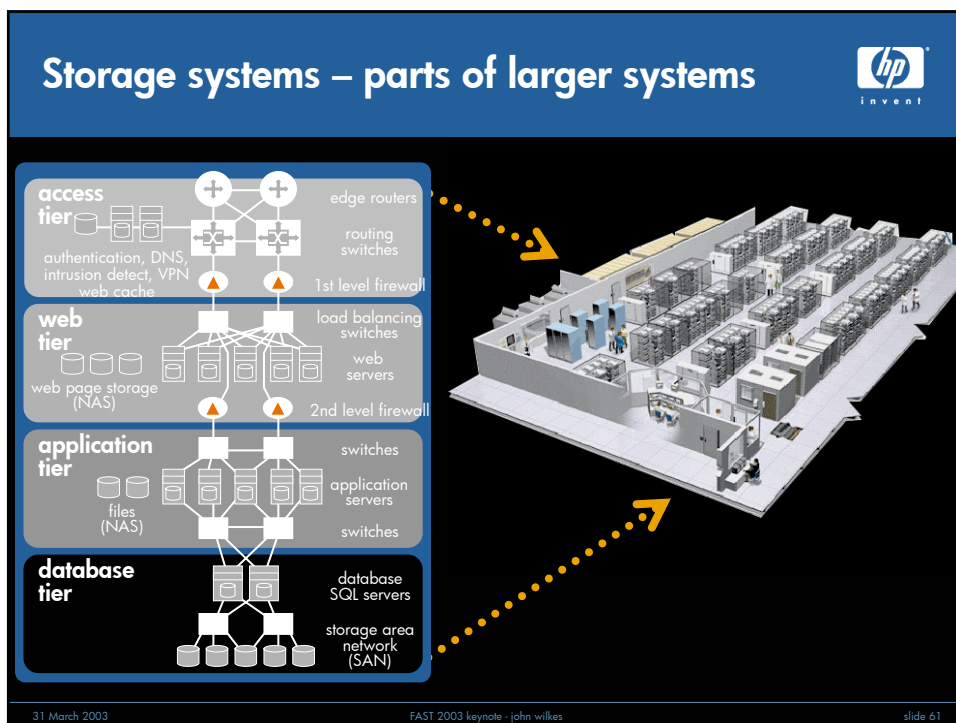
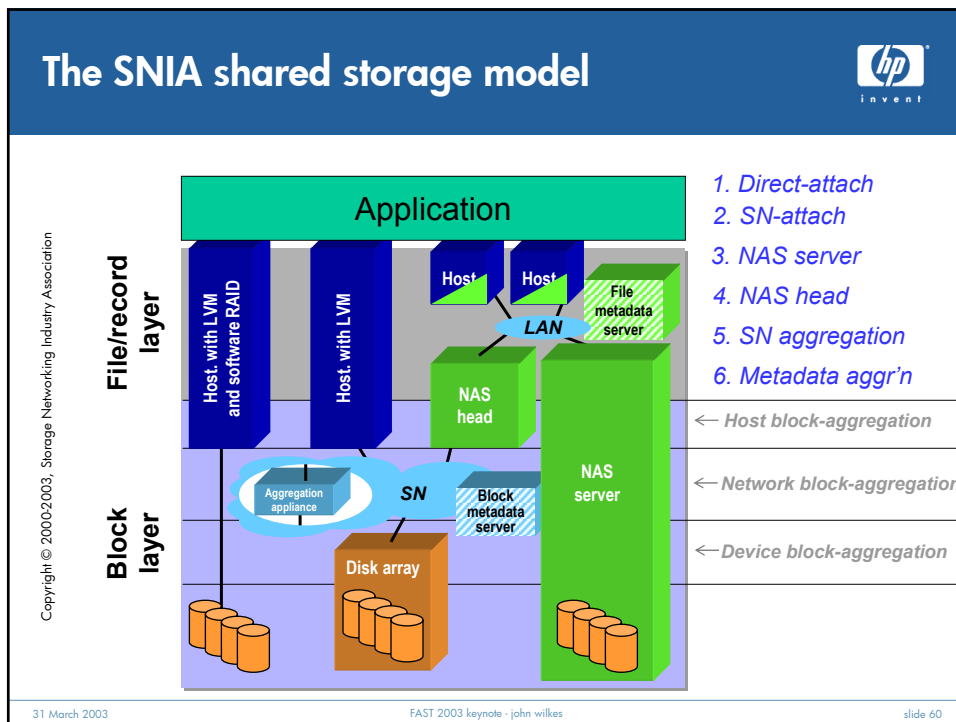
Storage domain

Services

- Discovery, monitoring
- Resource mgmt, configuration
- Security, billing
- Redundancy mgmt (backup, ...)
- High availability (fail-over, ...)
- Capacity planning

Copyright © 2000-2003, Storage Networking Industry Association

31 March 2003 FAST 2003 keynote - john wilkes slide 59





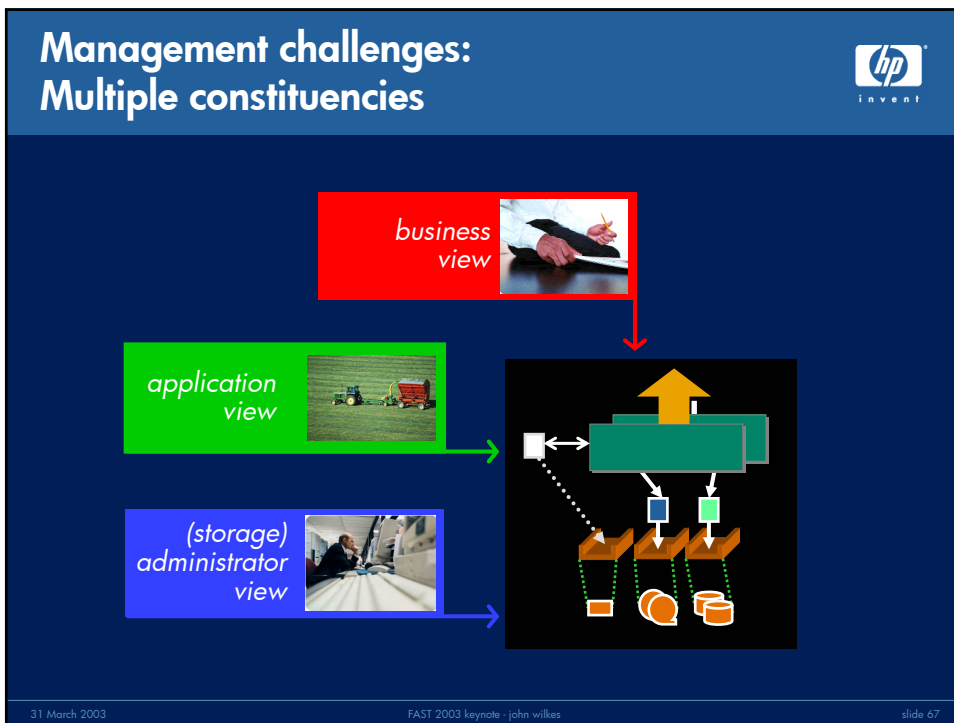
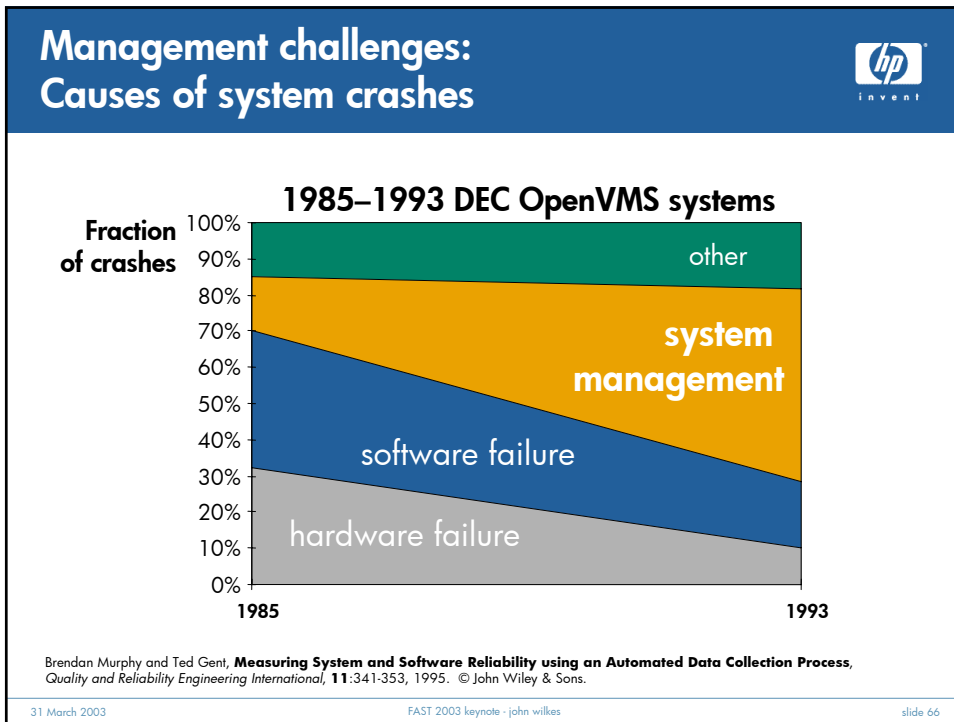
## Management challenges: Administration cost




- Rules of thumb:
  - 1980: 1 data administrator / 10GB
  - 2000: 1 data administrator / 5TB
- Problem:
  - 5TB ~ \$5k in a few years
  - admin cost >> storage cost!
- Conclusion: need to automate **all** storage administration tasks



Adapted from *Storage bricks have arrived*  
Jim Gray, Microsoft Research, FAST 2002



## Management challenges: some of the things to address




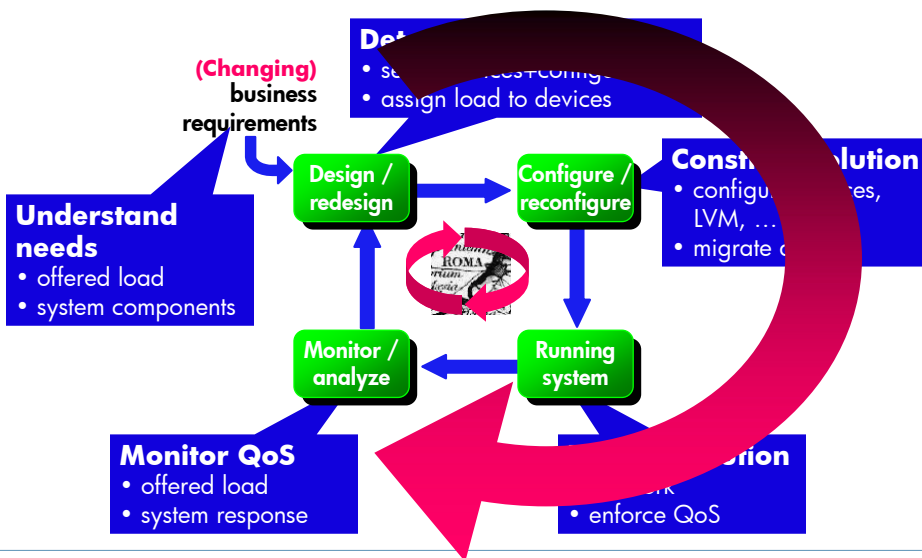
- Physical infrastructure
  - capacity planning
  - discovery, installation
  - allocation, qualification
  - configuration
- Logical configuration
  - data placement, security
  - volume/device virtualization
- QoS enforcement (runtime)
  - security, performance
  - data protection, recovery
  - consistency/coherency
- Monitoring
  - reporting
  - billing



31 March 2003
FAST 2003 keynote - john wilkes
slide 68

## Management challenges: full automation





31 March 2003
FAST 2003 keynote - john wilkes
slide 69



## Management challenges: the design step



To achieve **complete** automation:

- Specify what's wanted (goals), not how to achieve it (implementation)
- Management system must choose what to do, not people
- Human oversight + feedback to correct/refine choices

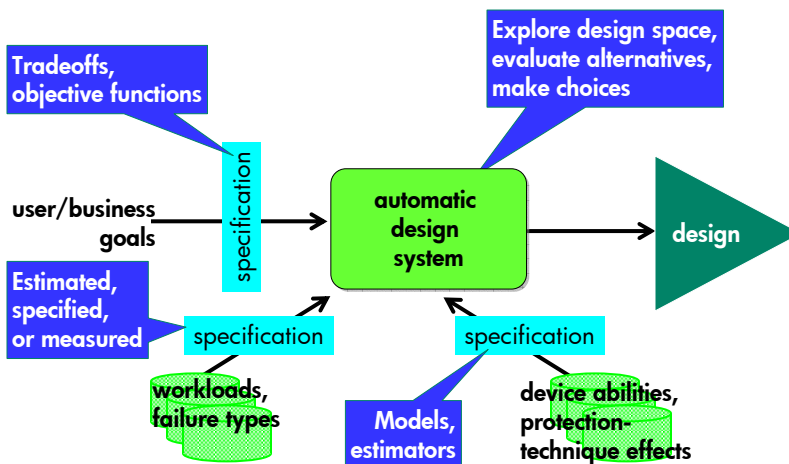


31 March 2003

FAST 2003 keynote - john wilkes

slide 70

## Management challenges: the design step




31 March 2003


FAST 2003 keynote - john wilkes

slide 71

## Management challenges: the design step




- There are many techniques available to us
  - which one to use?
  - in what circumstances?
  - with what settings?
  - are they working?
- **Potential provocative idea:** ~~no new techniques~~ until we can define these properties?

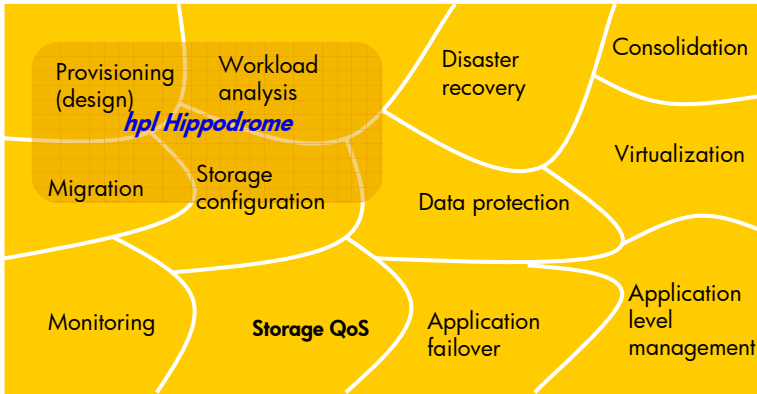


31 March 2003 FAST 2003 keynote - john wilkes slide 72

## Management challenges: future work




**There's plenty of scope for it!**




31 March 2003 FAST 2003 keynote - john wilkes slide 73



From some of the parts ...  
... to the sum of the parts



**"some assembly required"**




**data service-driven storage systems**

31 March 2003 FAST 2003 keynote - john wilkes slide 75

The slide features a blue header with the text "From some of the parts ... to the sum of the parts" and the HP logo. Below the header, on the left, is a photograph of scattered yellow puzzle pieces against a blue sky, with the text "some assembly required" in quotes. A large blue arrow points from this image to a photograph on the right showing a completed yellow puzzle structure. Below the puzzle structure is the text "data service-driven storage systems". At the bottom of the slide, there are three small text elements: "31 March 2003", "FAST 2003 keynote - john wilkes", and "slide 75".

## Data services



**Anywhere, anytime access to data**

QoS-driven:


- affordable
- flexible
- predictable
- reliable

In a rapidly changing world,  
at scales that dwarf the desktop,  
while leaving people in control



31 March 2003 FAST 2003 keynote - john wilkes slide 76


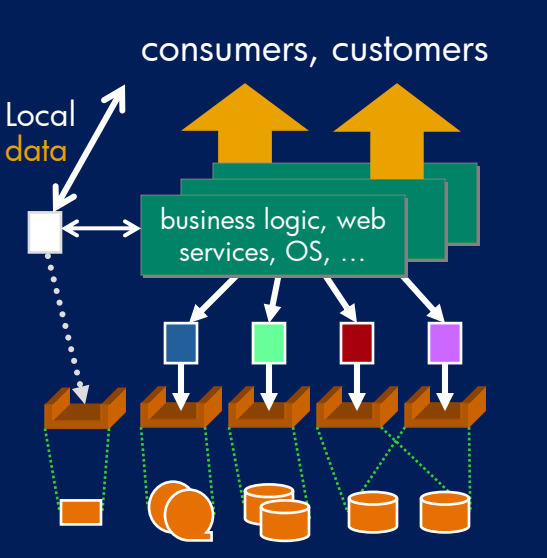
## Data services-driven storage systems



consumers, customers

Local data

business logic, web services, OS, ...



31 March 2003 FAST 2003 keynote - john wilkes slide 77

## Data services



Rising **system complexity** +  
rising **abilities** +  
rising **expectations**

Solution:


- define **data QoS** needs
- use **storage QoS** abilities
- **automate** storage + data management

Our target should be:  
**data services**



31 March 2003 FAST 2003 keynote - john wilkes slide 78


## Data services



**It's going to be an exciting ride, at all levels:**

- storage devices
- storage networking
- storage management
- data management

**Welcome aboard!**



31 March 2003 FAST 2003 keynote - john wilkes slide 79

## Data services – from data to containers



**thank you!**

john.wilkes@hp.com  
<http://www.hpl.hp.com/research/ssp>