WOSP ‘98, Santa Fe, NM, 12-16 October 1998

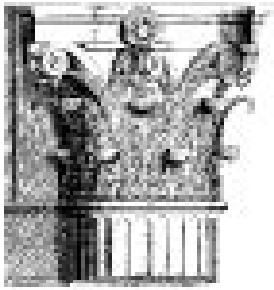
Capacity planning with phased workloads

Arif Merchant

Storage Systems Program
Computer Systems Laboratory
Hewlett-Packard Laboratories, Palo Alto, CA

Joint work with E. Borowsky, R. Golding, P. Jacobson, L. Schreier, M. Spasojevic and John Wilkes

HEWLETT PACKARD

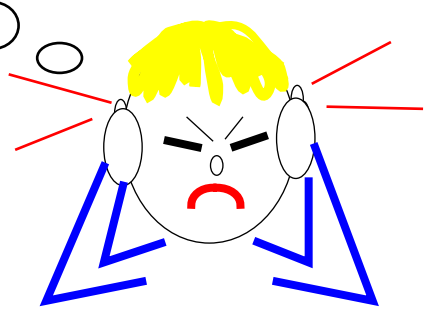


Attribute-managed storage

A day in the life of a System Administrator

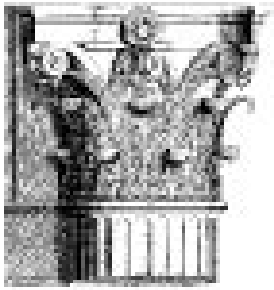
Need more capacity.
Need better performance.
Need high availability.
Must rebalance the load.
Must add devices.
UGH!...
my head hurts!

Quality of service guarantees.
Network attached storage.
More demanding applications.
AAAGH!...
Brain exploding!



Headache today?

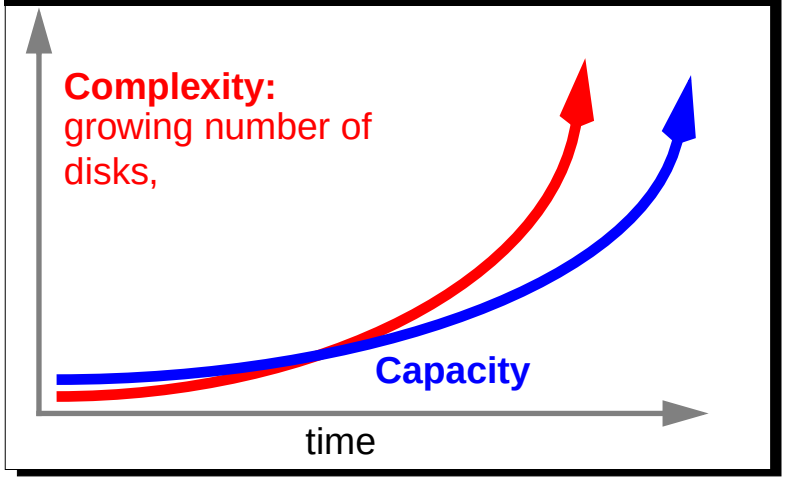
Migraine tomorrow!!!



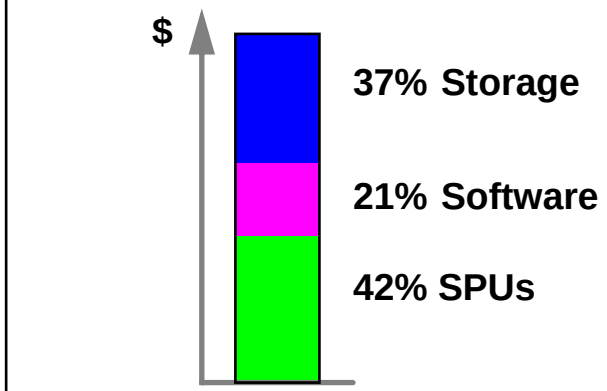
Attribute-managed storage

Motivation

Growing complexity of storage systems

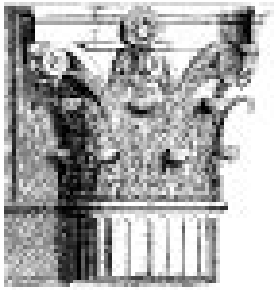


HP 300GB TPC-D benchmark list price



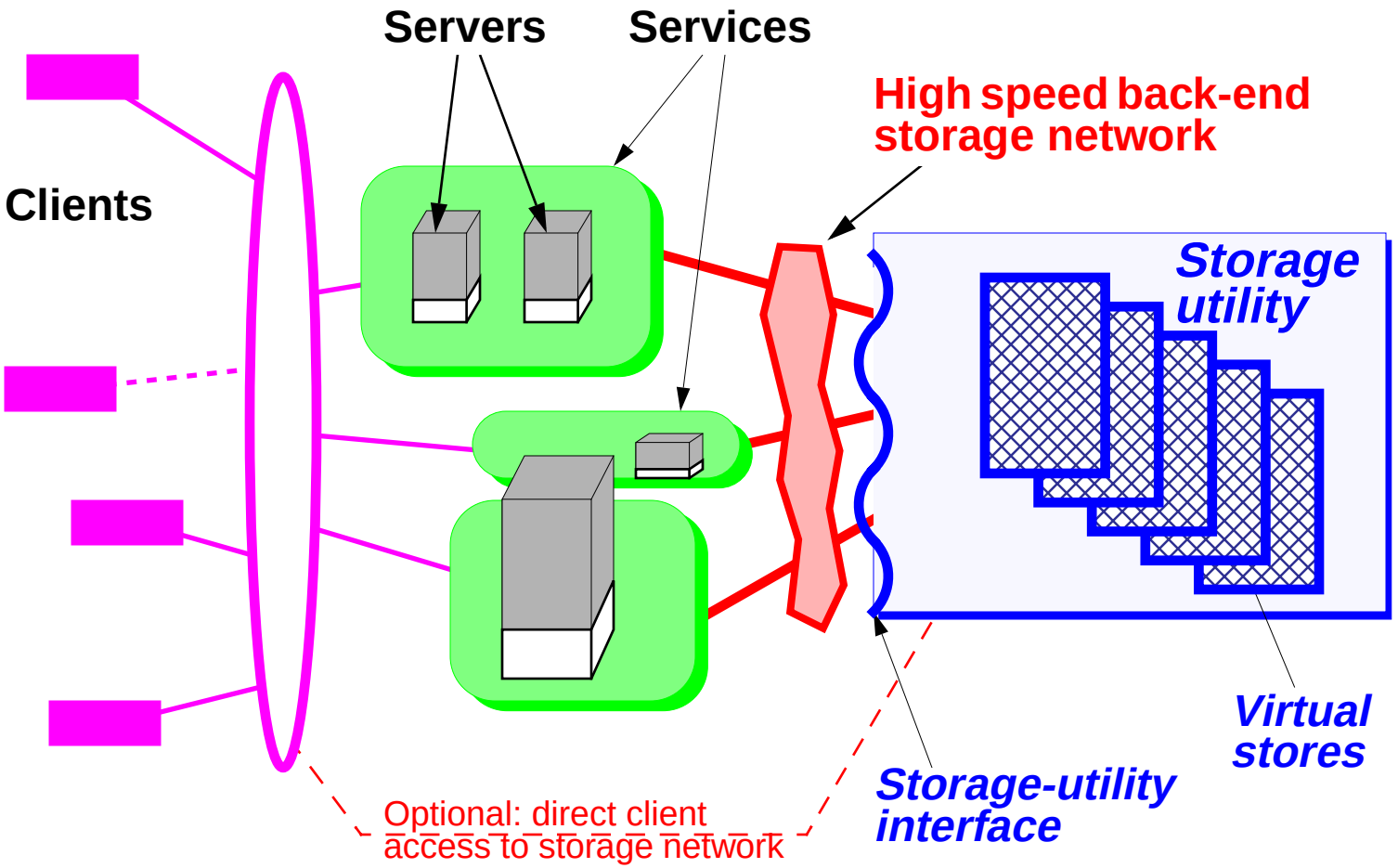
Growing cost of ownership for storage systems

HEWLETT PACKARD

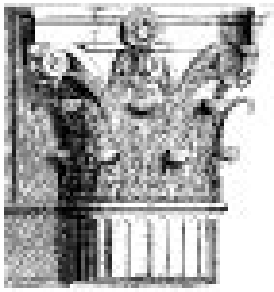


Attribute-managed storage

Opportunity

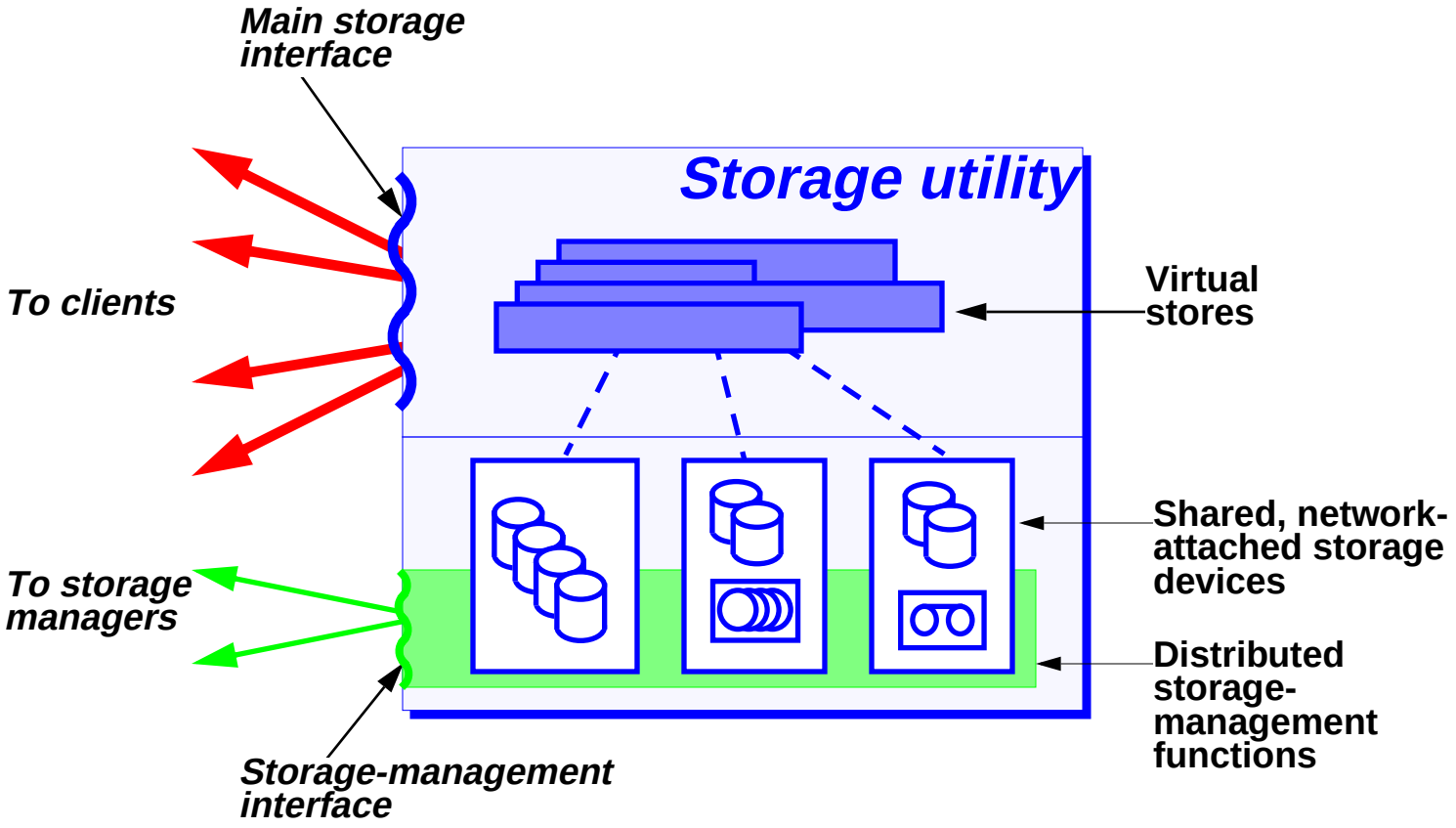


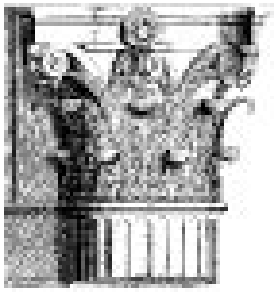
HEWLETT PACKARD



Attribute-managed storage

A closer look





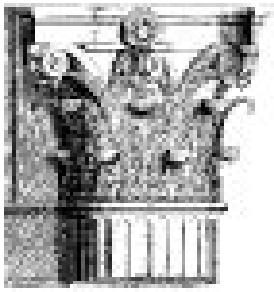
Attribute-managed storage

The goal

Say what you want not how to do it!

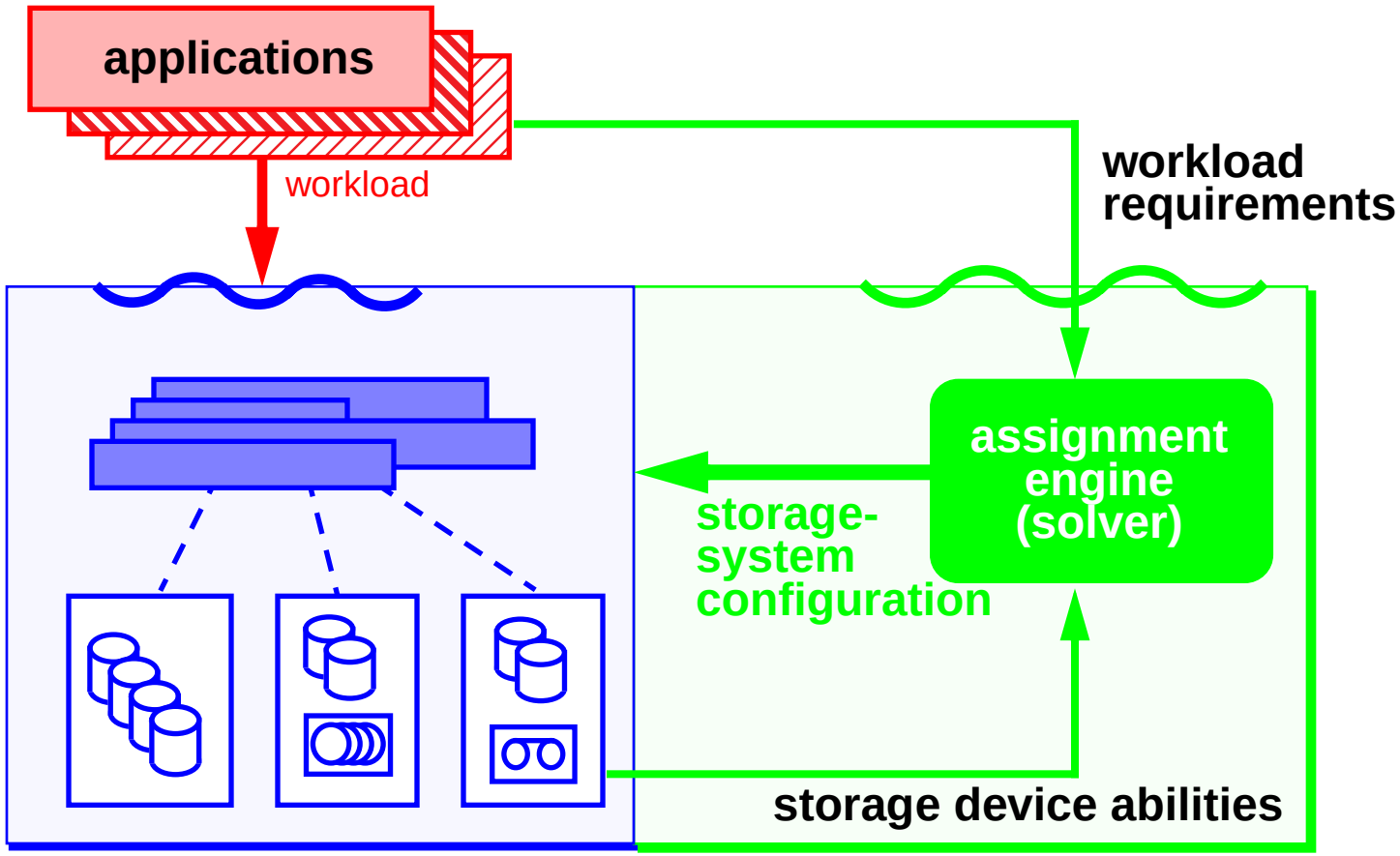
RAID 3 data layout, across 5 of the disks on disk array F, using 64KB stripe size, 3MB dedicated buffer cache with 128KB sequential readahead buffer, delayed write-back with 1MB NVRAM buffer and max 10s residency time, dual 256Kb/s links via host interfaces 12.4.3 and 16.0.4, 1Gb/s trunk links between FibreChannel switches A-3 and B-1, …

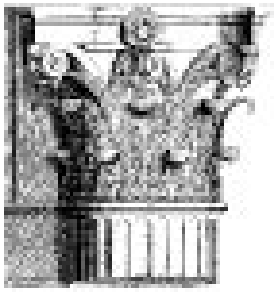
- business-critical availability
- 100 IOs/sec
- 200ms response time



Attribute-managed storage

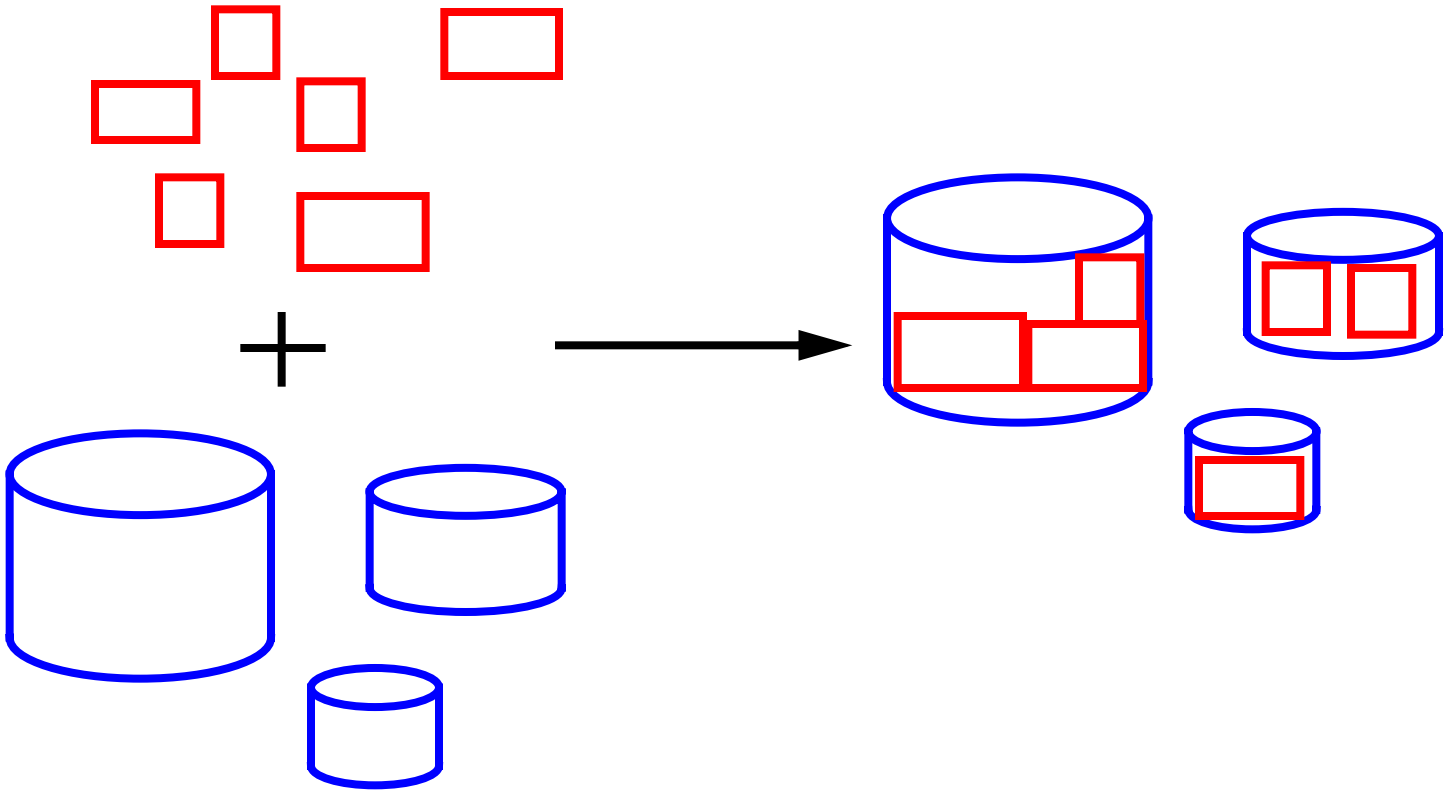
The mechanism





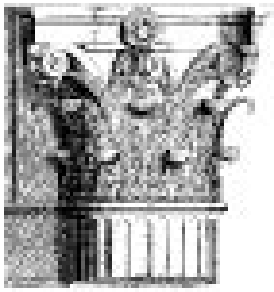
Attribute-managed storage

The assignment problem

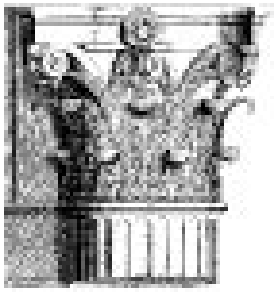


Constraints

Does it fit?



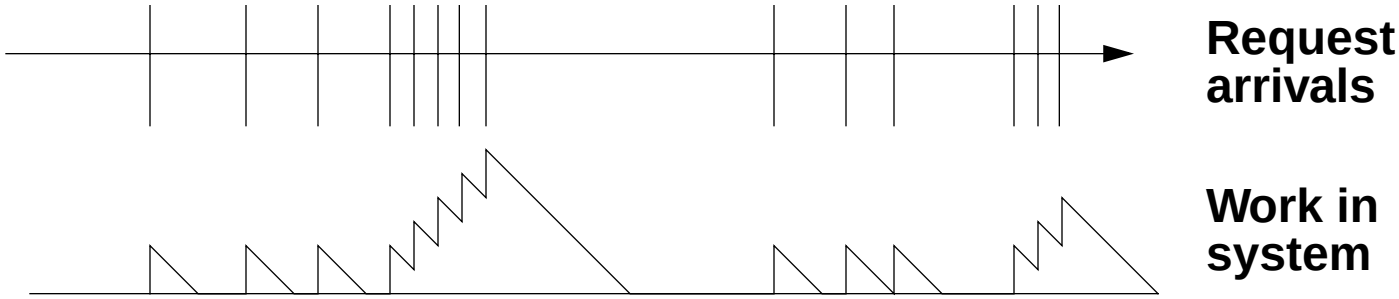
- **Capacity constraints**
  - Is there enough space?
- **Availability constraints**
  - Is it up often enough?
- **Performance constraints**
  - Is response time adequate? E.g.: Are 95% of requests satisfied within 0.2 sec?



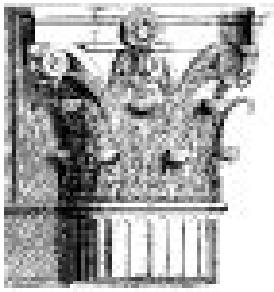
Short Term Utilization

Intuition

Queues form in stable system because of variation in workload arrival rate.



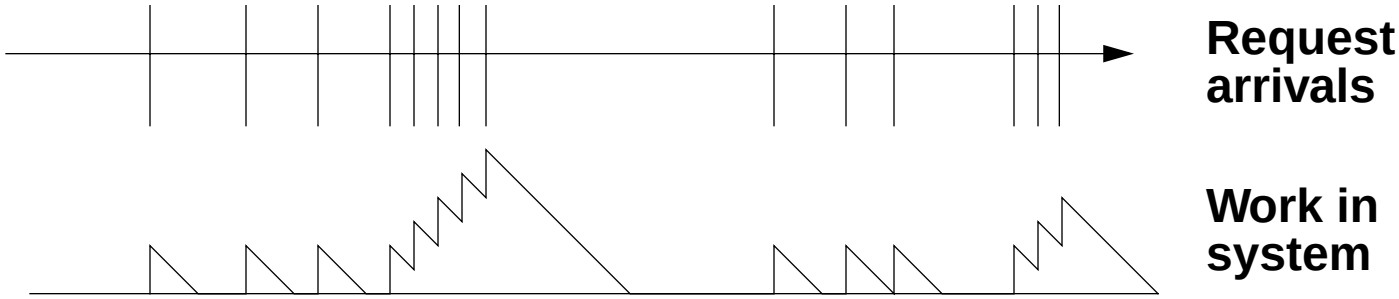
Queueing delays can be controlled by controlling variability in work arrival rate.



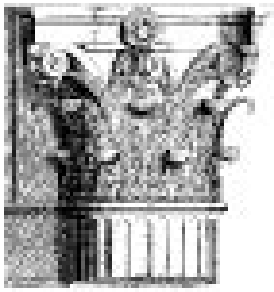
Short Term Utilization

A theorem

If the work arriving in every period of length T is such that the device can do it in T seconds, then the response time is always less than T seconds.



- **Setting T= maximum response time allowed meets requirements.**
- **But ... this requirement is too strict.**

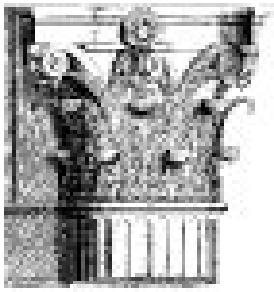


Short Term Utilization

An approximation

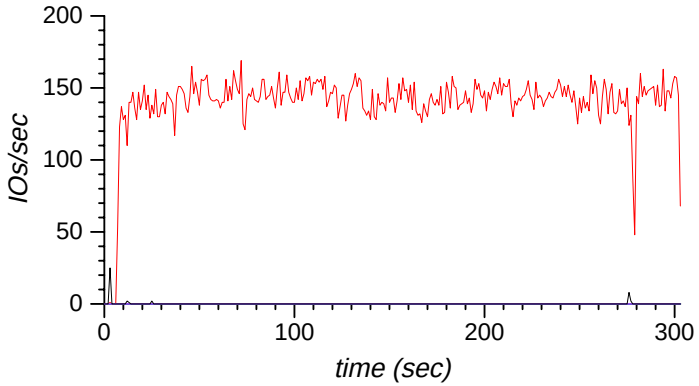
Pr{Work arriving in T < what device can do in T} > p =>
Pr{Response time < T} > p

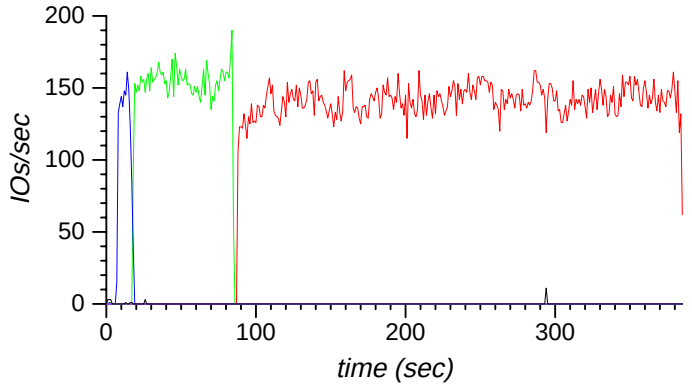
- Translates bound on response time tail into a bound on tail of Work(T)
- Approximation is exact for p=1
- Distribution of Work arriving in time T frequently easy to calculate or approximate for simple workloads.

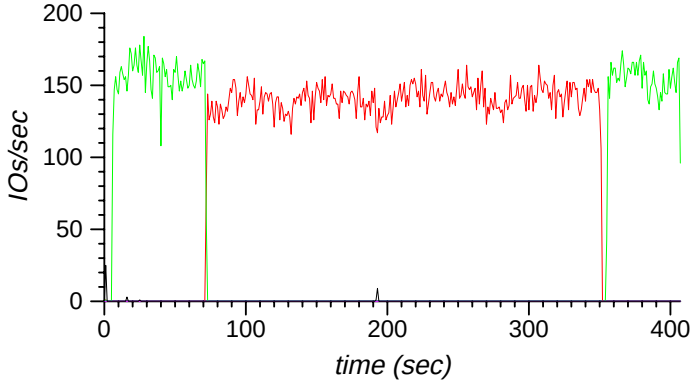


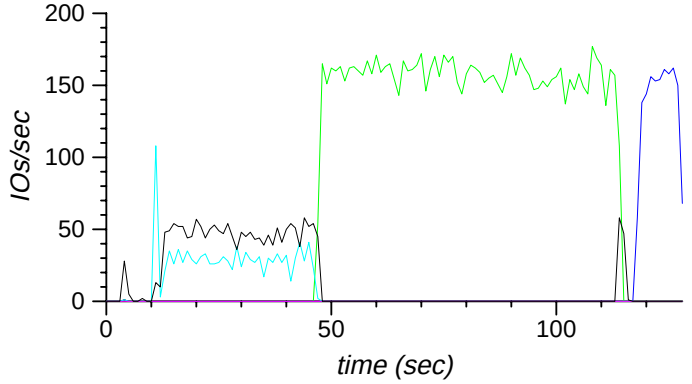
Workload Characterization

TPC-D workload traces: application phases

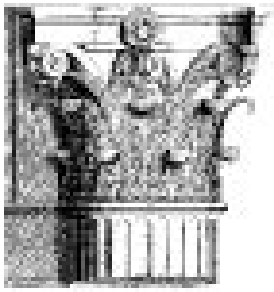








HEWLETT PACKARD



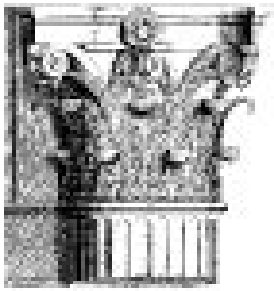
Workload characterization

Phased correlated model

Each workload is modeled as a ON-OFF Poisson process

- Parameters: ON time average, OFF time average, IO rate during ON period
- Correlation between workloads:

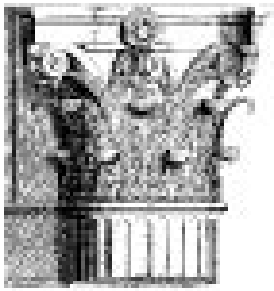
  pij = Pr{Aj is ON when Ai comes ON}



Phasing and Short term utilization

Combining forces

- Response times increase only when some workload goes ON
- Sufficient to test response time bounds only at the times workloads change state from OFF to ON
- Workload distribution is easy to estimate given a workload just went ON.

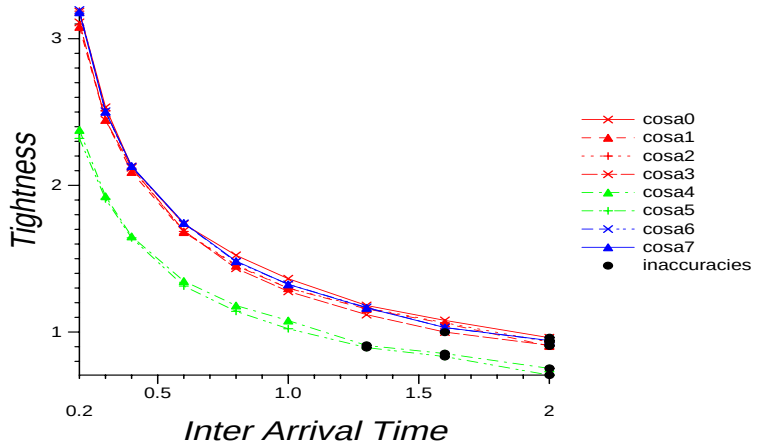


Validation and testing

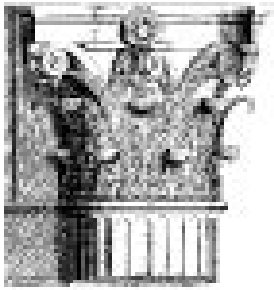
Tasting the stew

Compared simulation and modelling results

- Baseline case: 8 streams, correlated sets of 4,2, 2. All predictions were correct.
- Checking tightness of predictions - are the bounds optimistic (wrong) or pessimistic?

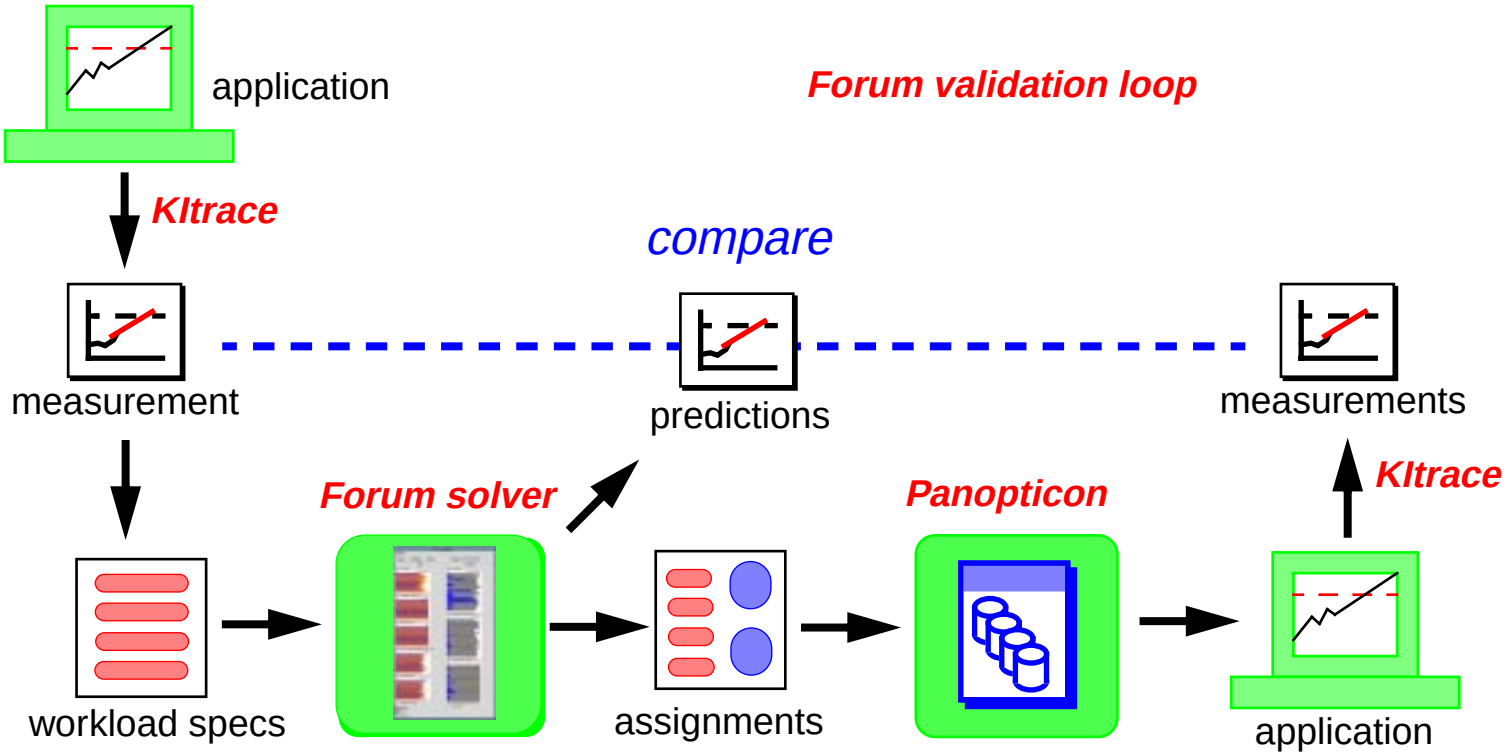


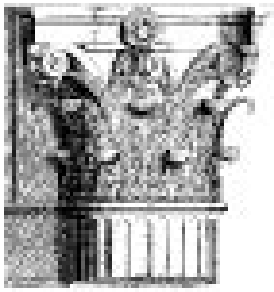
HEWLETT PACKARD



Validation and testing

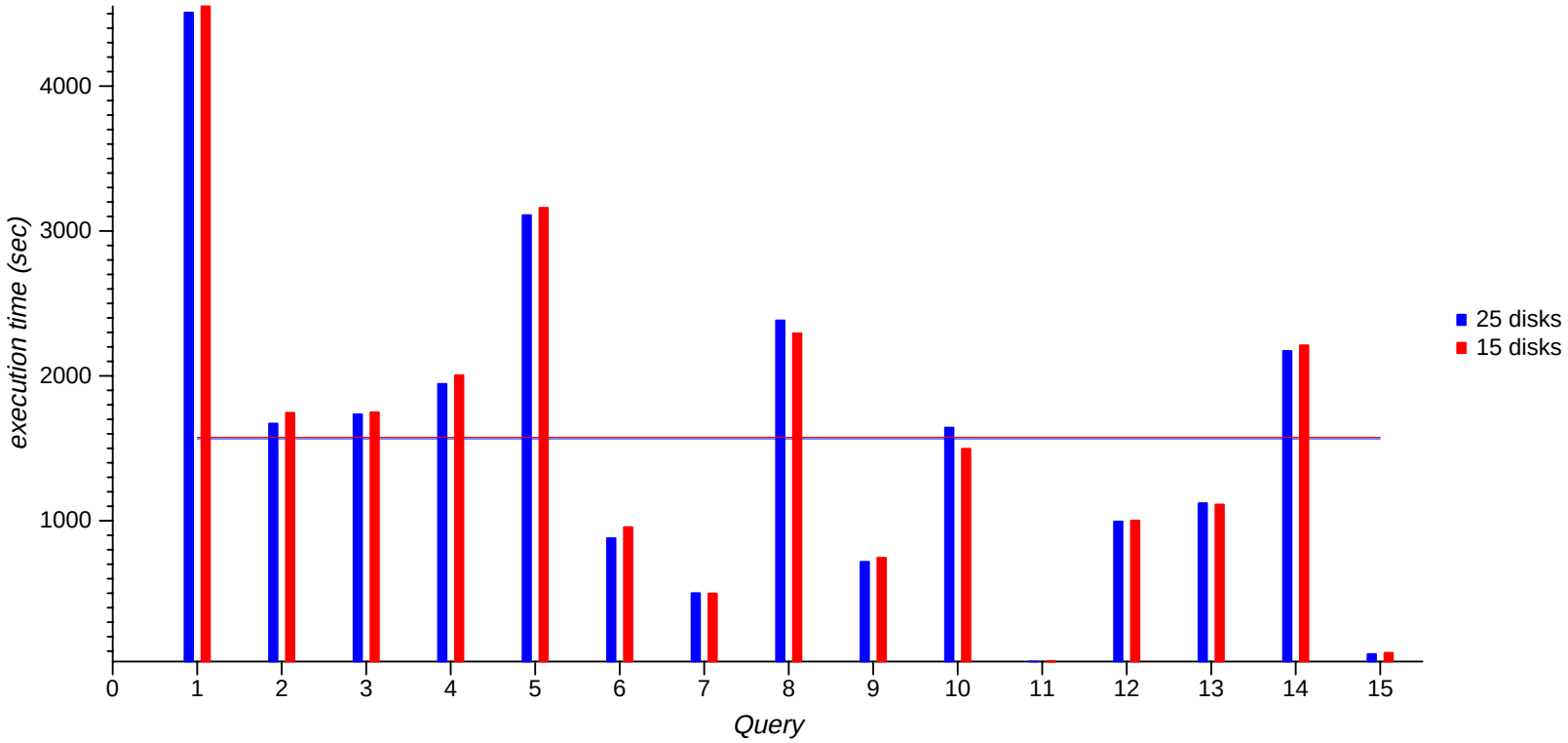
The validation loop





Validation and testing

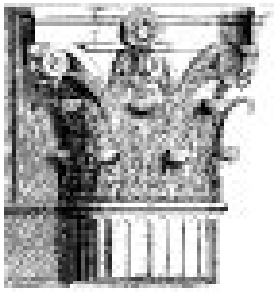
The pudding



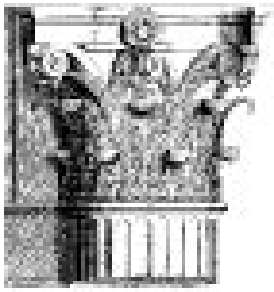
Query execution times: 25 vs. 15 disks

Capacity planning

What next?



- Better device models
- Better workload models
- Fault-tolerant on-line management



Attribute-managed storage

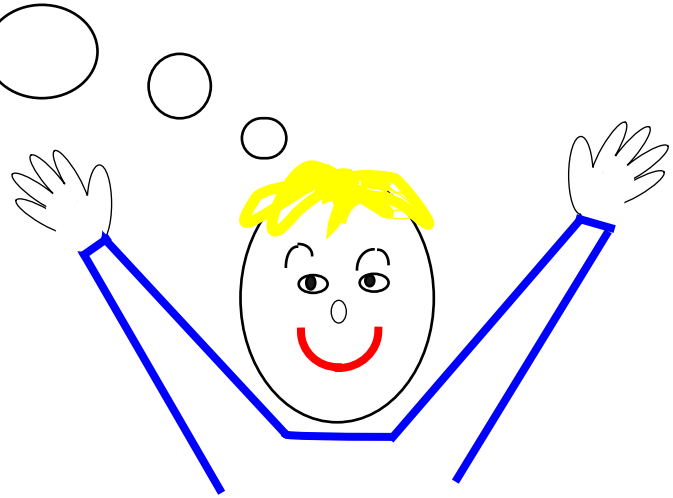
The future

Need guaranteed quality of service?

Storage distributed across the network?

Continually changing workload?

NO PROBLEM!



http://www.hpl.hp.com/SSP