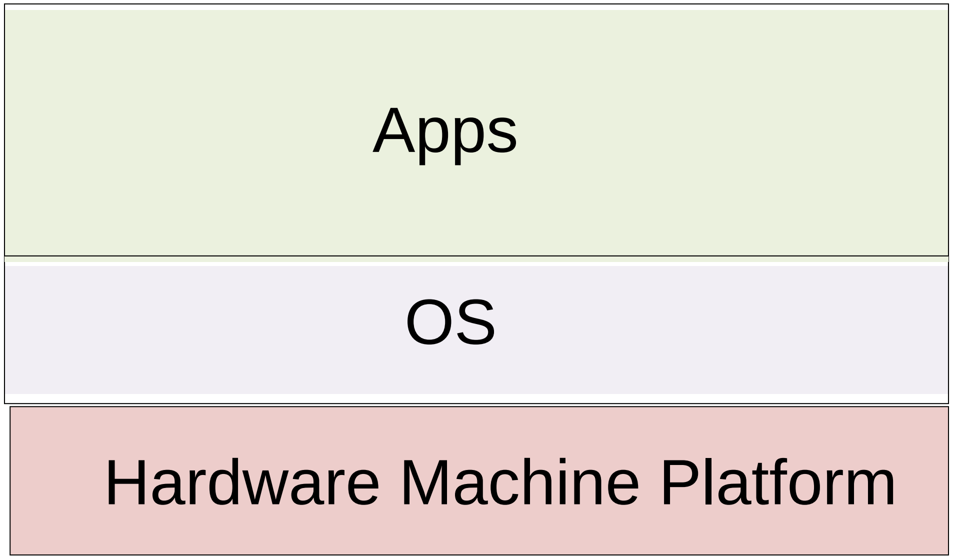
# EC2 demystification, server power efficiency, disk drive reliability

CSE 490h, Autumn 2008

# There's no magic to an OS

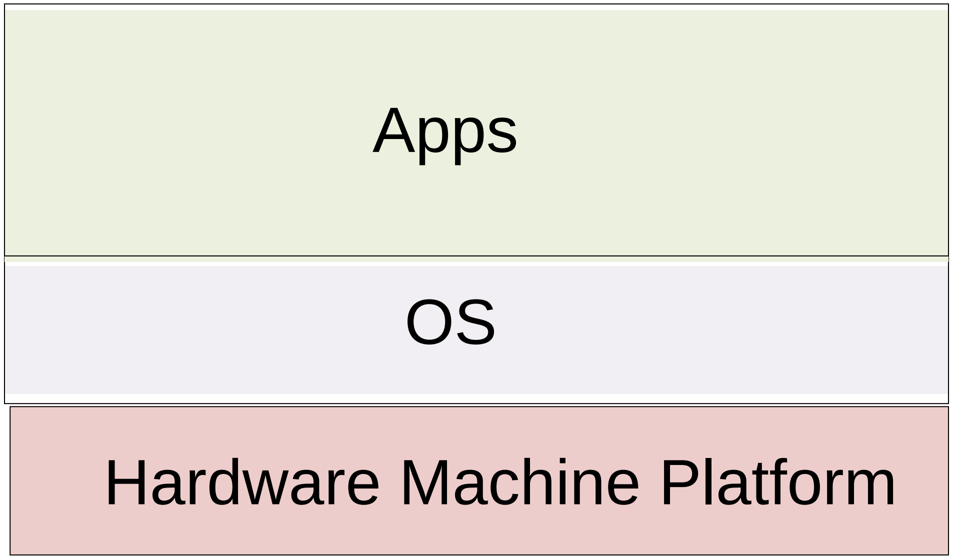


Apps

OS

Hardware Machine Platform

- How does an app do a file write?



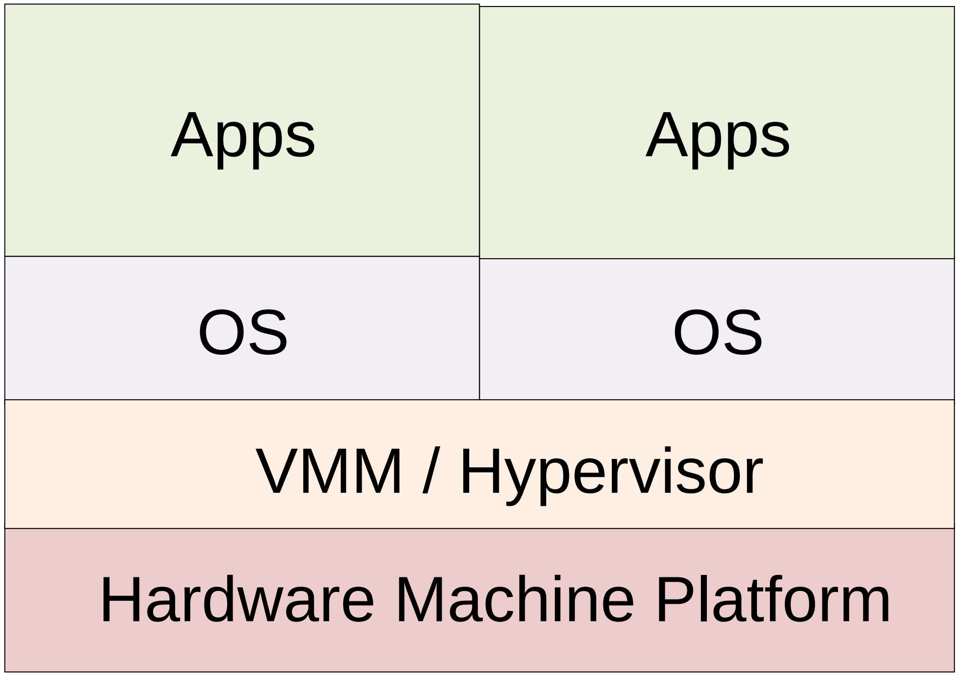
Apps

OS

Hardware Machine Platform

- How does an app do a file write?
- What happens if the app tries to cheat?

# There's no magic to a VMM



Apps

Apps

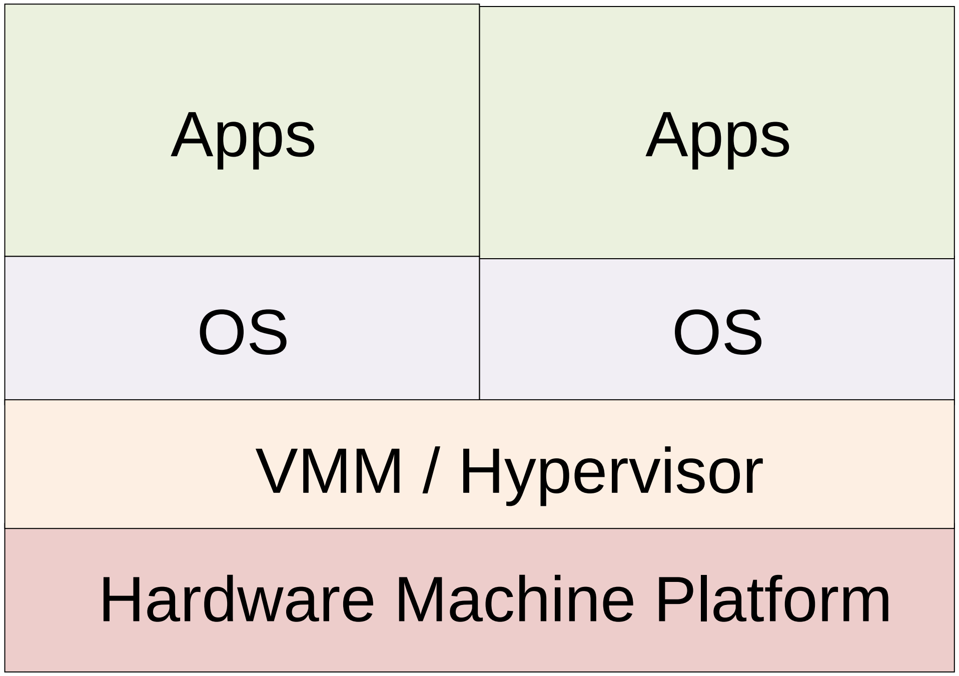
OS

OS

VMM / Hypervisor

Hardware Machine Platform

- How does an app do a file write?



Apps

Apps

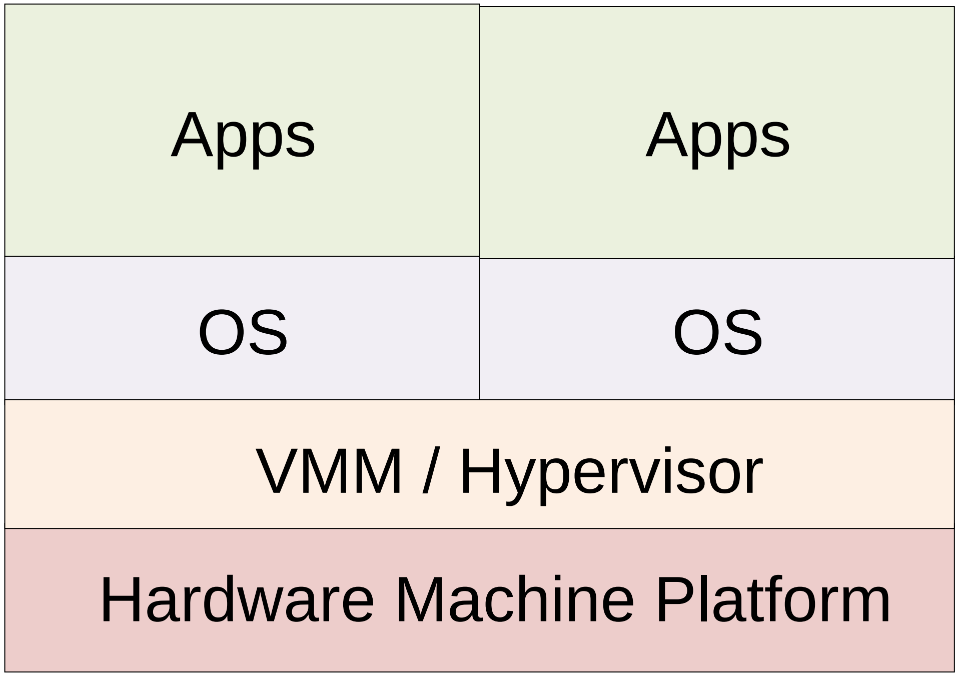
OS

OS

VMM / Hypervisor

Hardware Machine Platform

- How does an app do a file write?
- What happens when the guest OS attempts a disk write?



Apps

Apps

OS

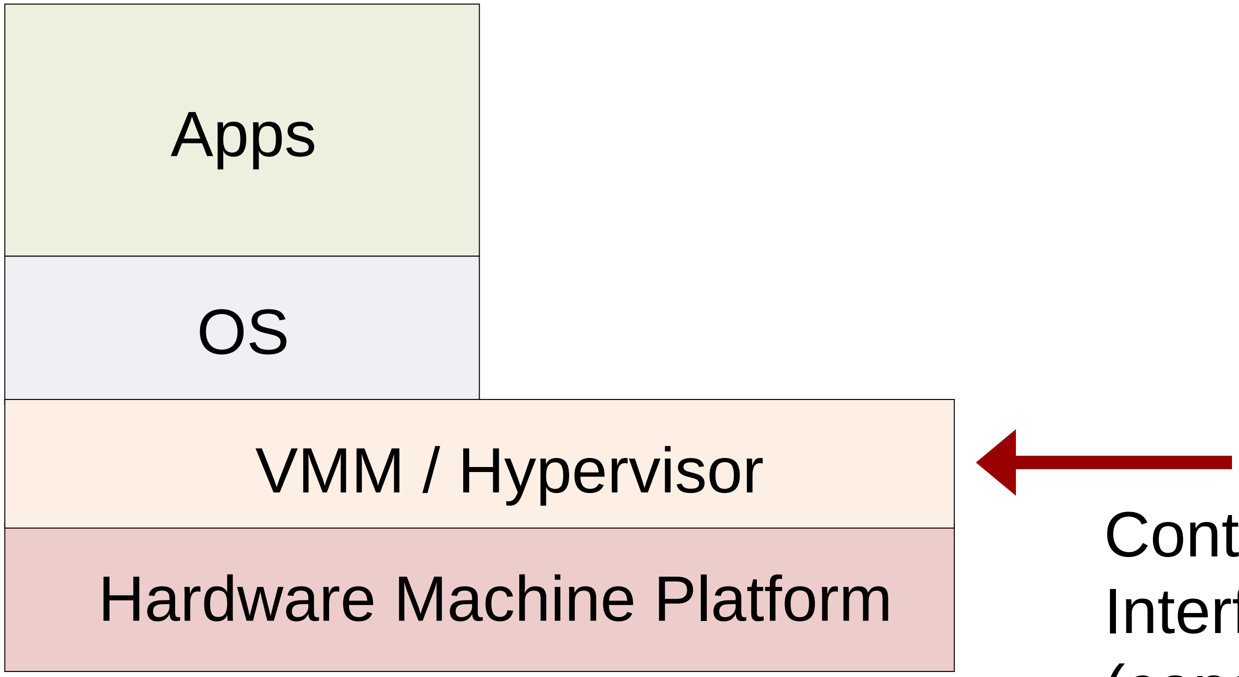
OS

VMM / Hypervisor

Hardware Machine Platform

- How does an app do a file write?
- What happens when the guest OS attempts a disk write?
- What happens if the app tries to cheat?

# There's no magic to creating a new VM



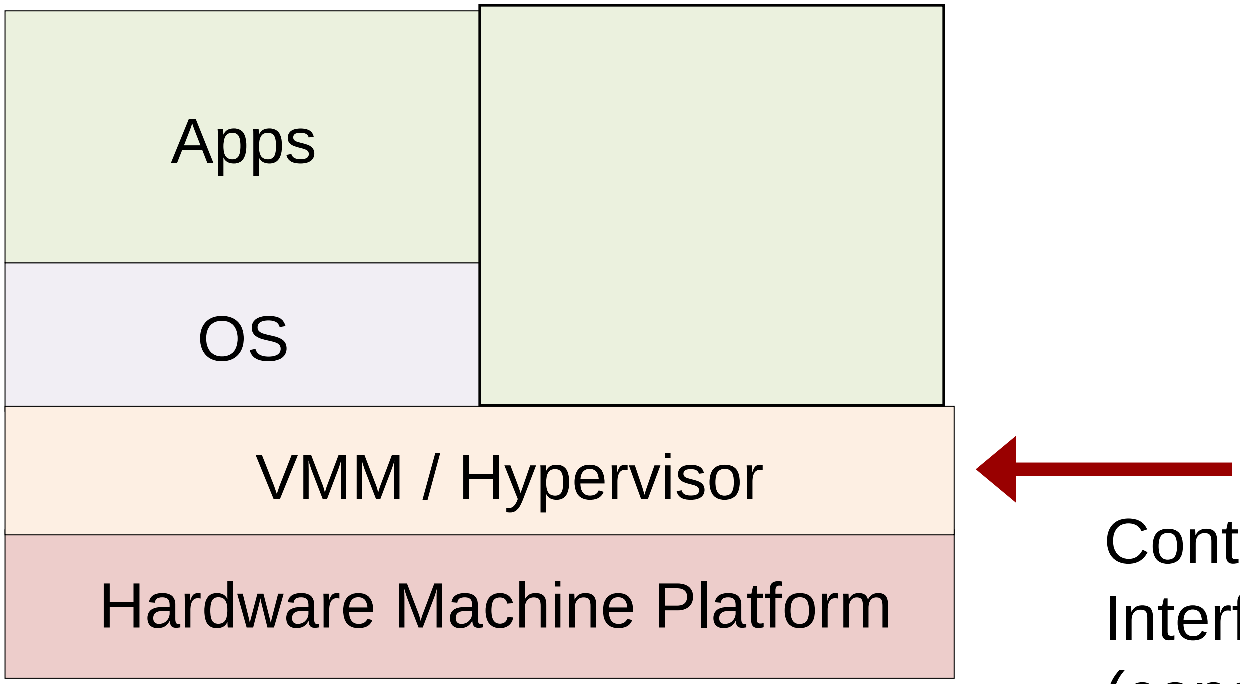
Apps

OS

VMM / Hypervisor

Hardware Machine Platform

Control
Interface
(console and
network)



Apps

OS

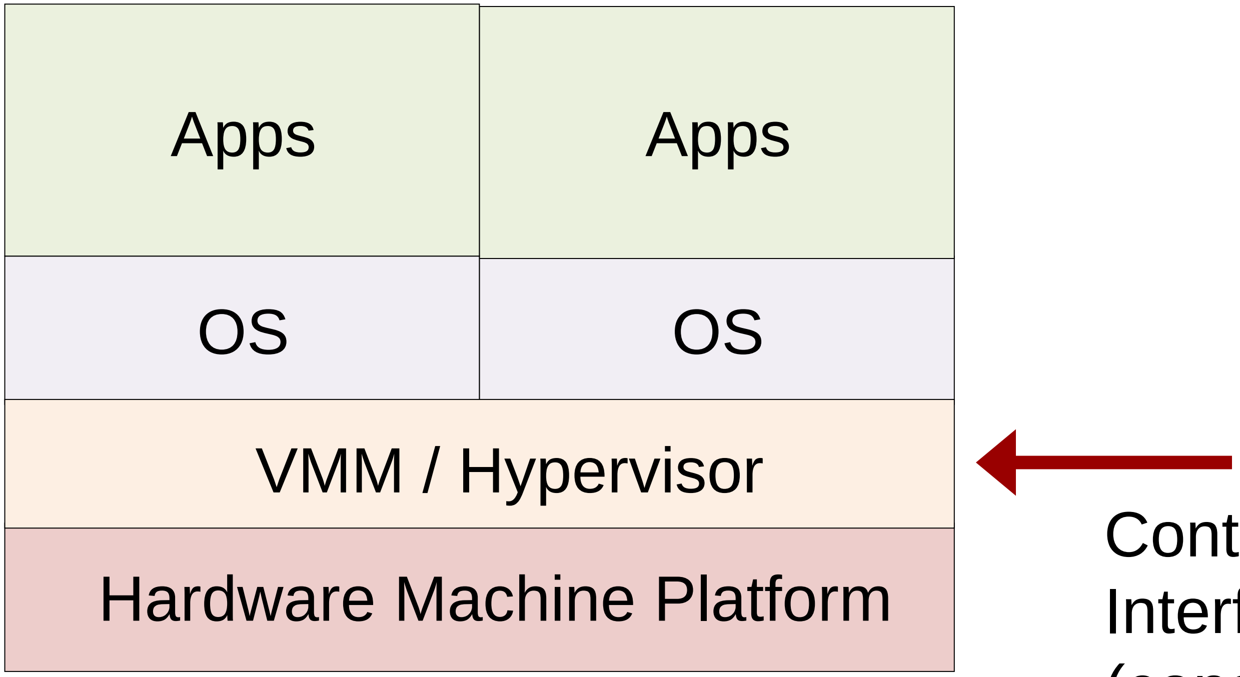
VMM / Hypervisor

Hardware Machine Platform

Control
Interface
(console and
network)

# There's no magic to creating a bootable system image

- Original UNIX file system
  - Boot block
    - can boot the system by loading from this block
  - Superblock
    - specifies boundaries of next 3 areas, and contains head of freelists of inodes and file blocks
  - i-node area
    - contains descriptors (i-nodes) for each file on the disk; all i-nodes are the same size; head of freelist is in the superblock
  - File contents area
    - fixed-size blocks; head of freelist is in the superblock
  - Swap area
    - holds processes that have been swapped out of memory
- And there are startup scripts for apps, etc.



Apps

Apps

OS

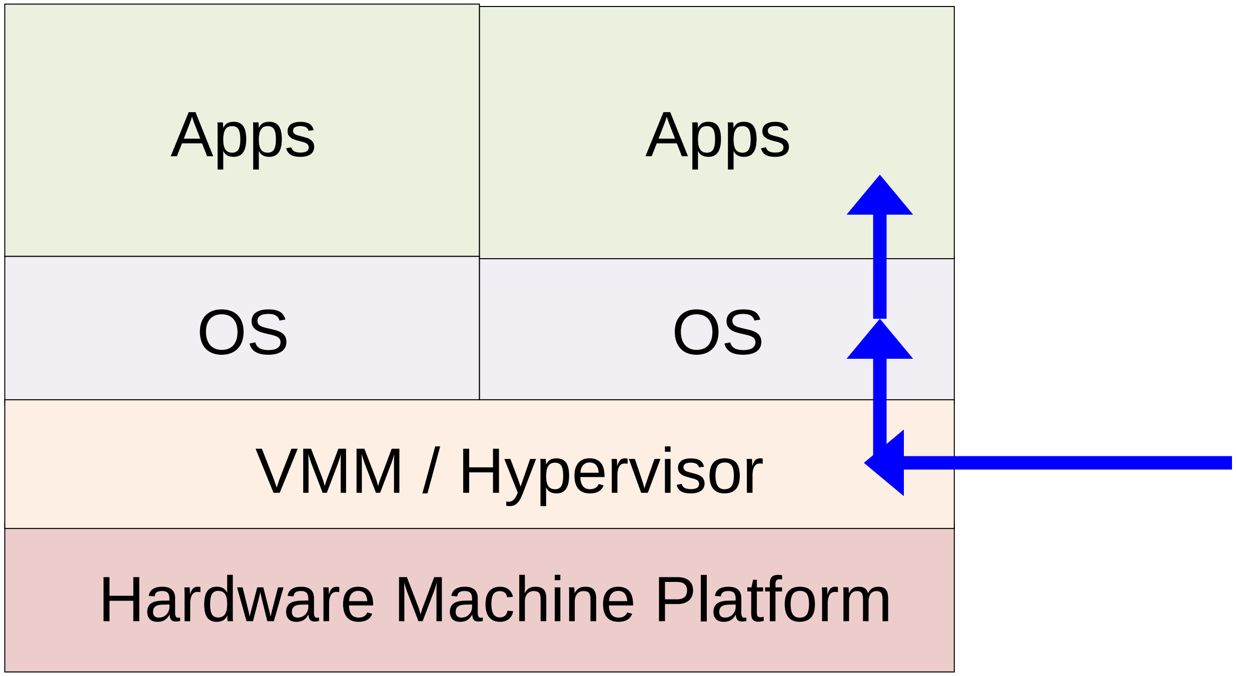
OS

VMM / Hypervisor

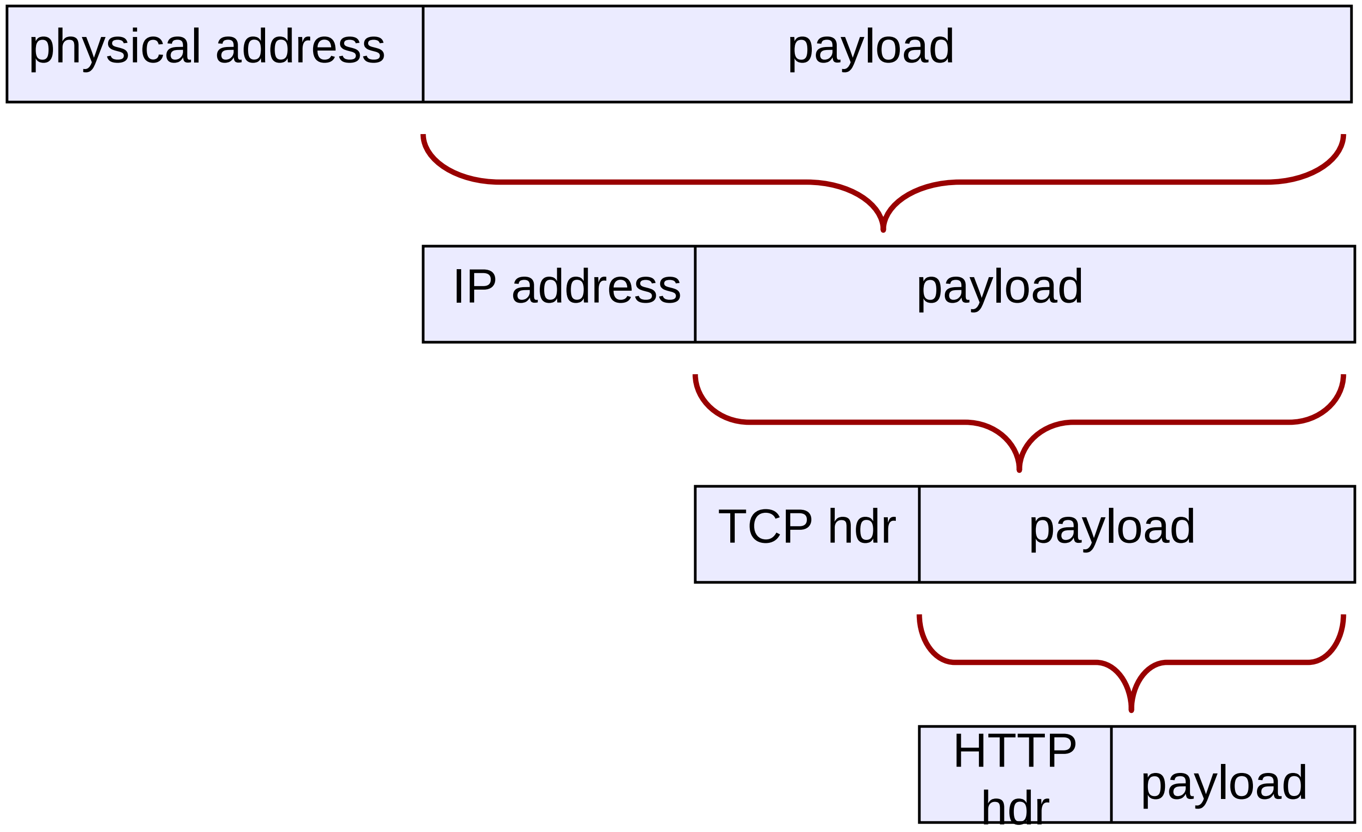
Hardware Machine Platform

Control
Interface
(console and
network)

# There's no magic to talking to your VM over the network



- Suppose your app was a webserver?



physical address | payload

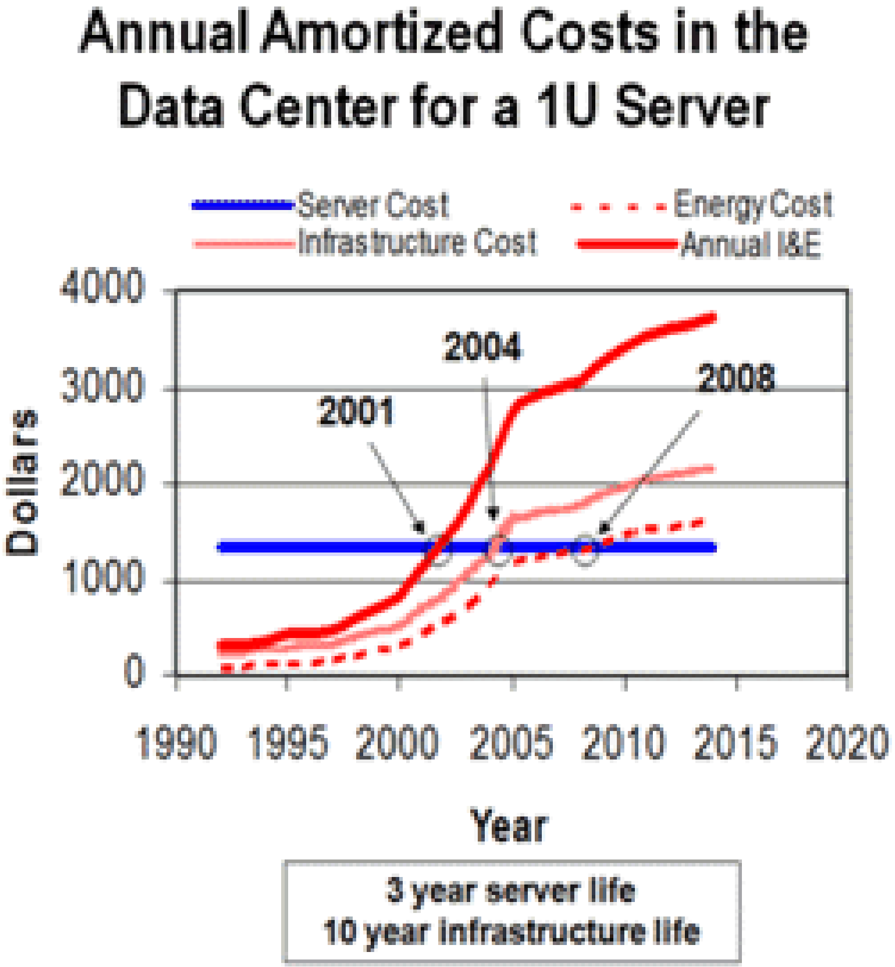
IP address | payload

TCP hdr | payload

HTTP hdr | payload

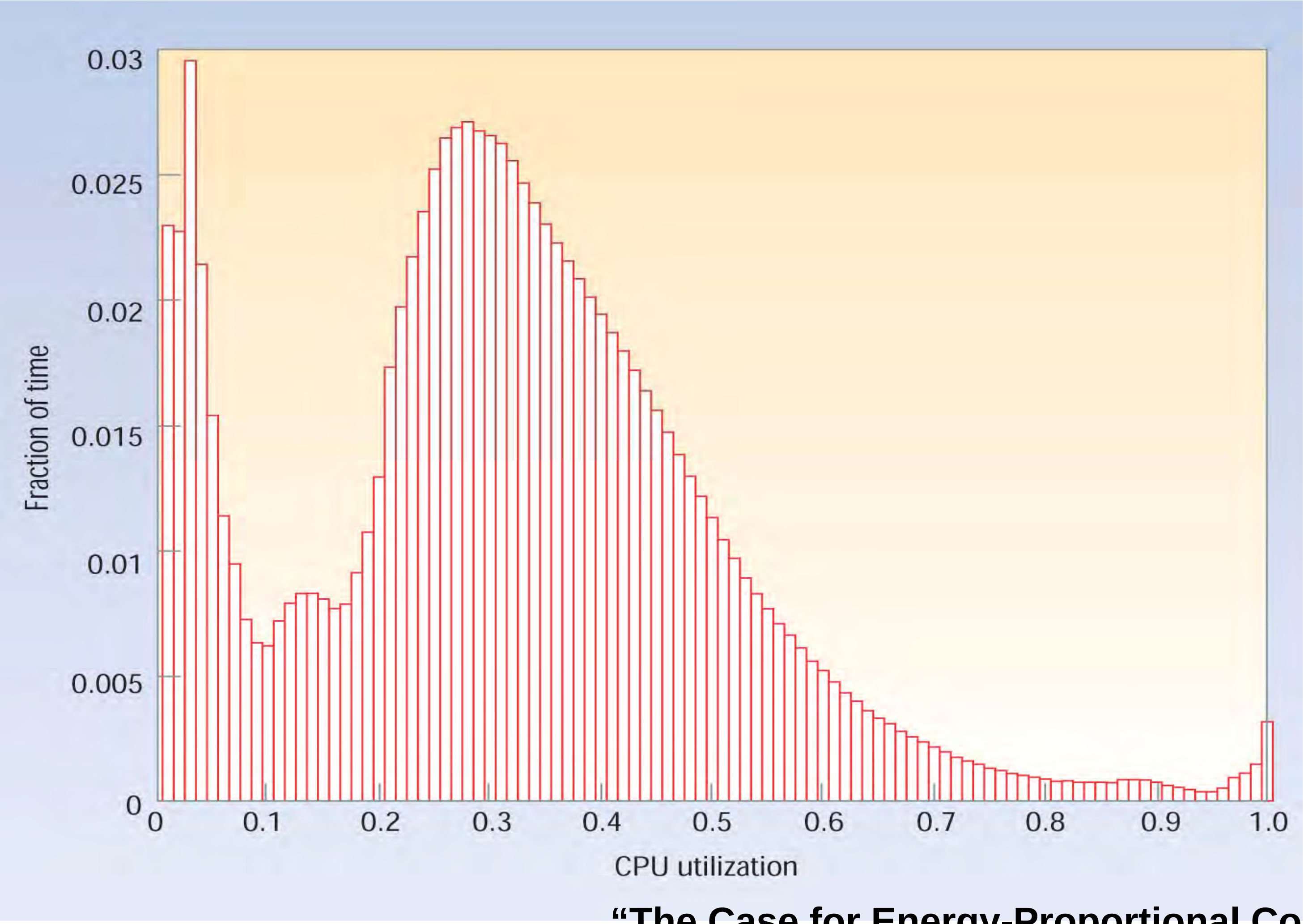
# Server power efficiency

- It matters

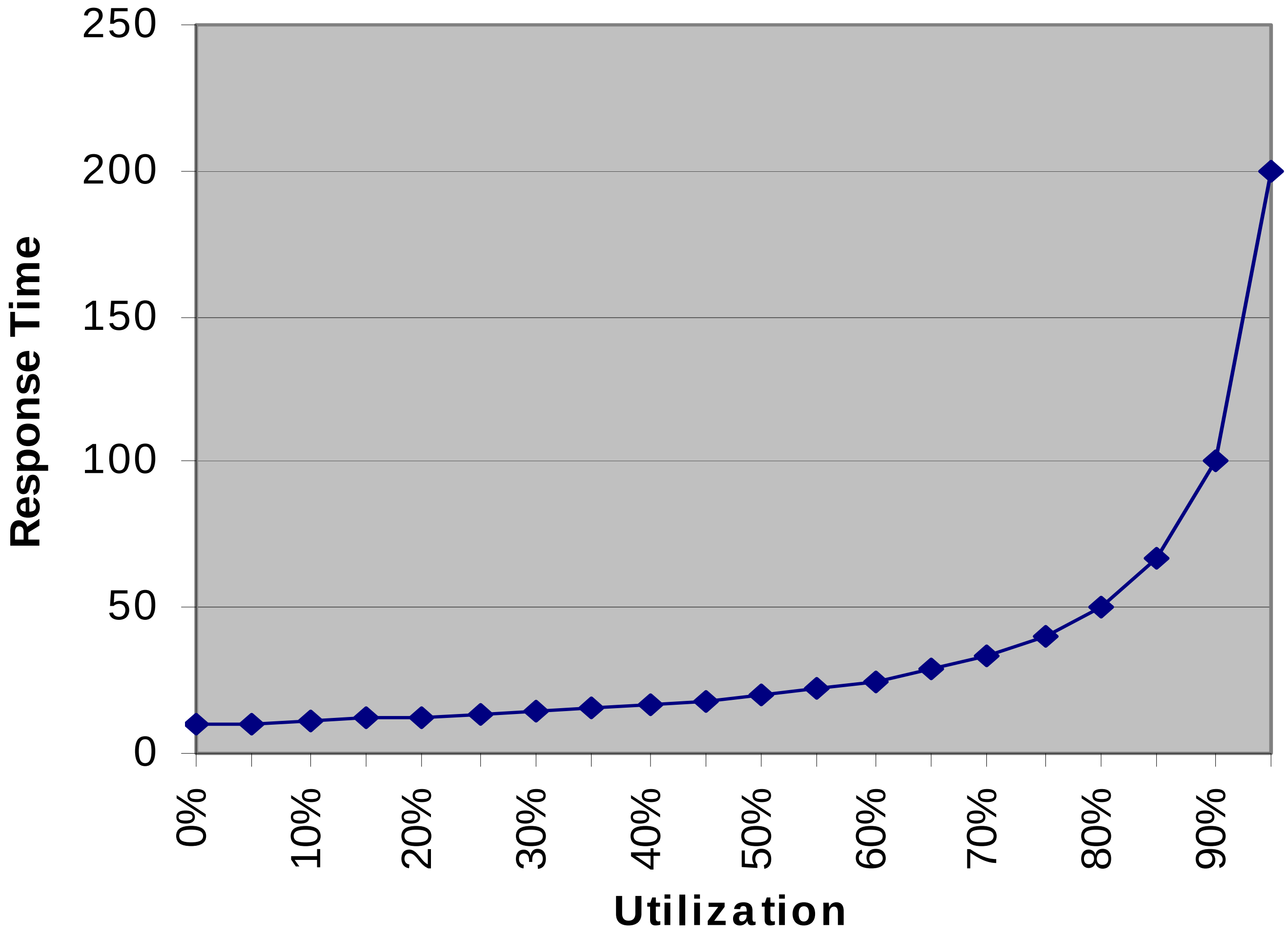


**http://www.electronics-cooling.com/articles/2007/feb/a3/**

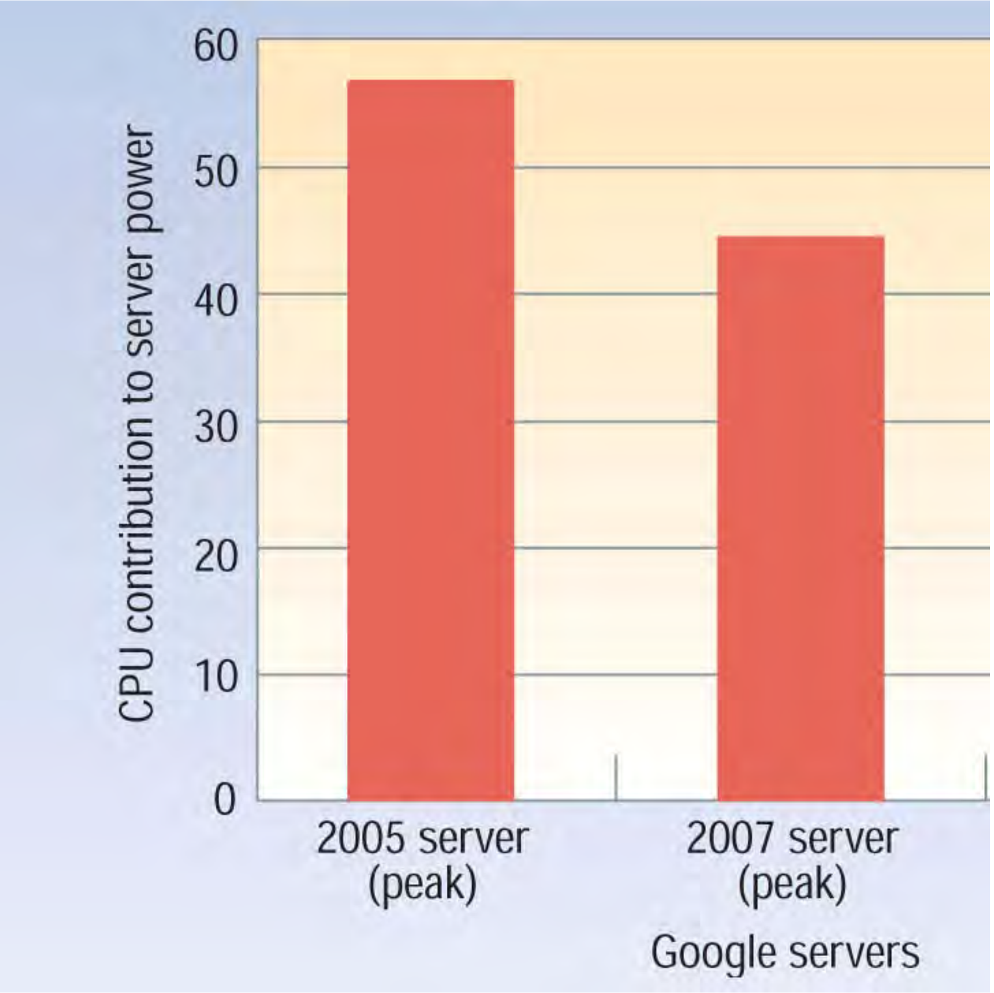
- Servers are typically operated at middling utilizations
- Necessary for performance reasons
  - Response time has a “knee” as utilization rises
- Terrible for energy efficiency
  - Only a 2:1 power consumption difference between low utilization and high utilization
- Very different than desktops
  - No one gave a rip about power consumption until recently
- Very different than laptops
  - Operate at peak or at idle, seldom in the middle



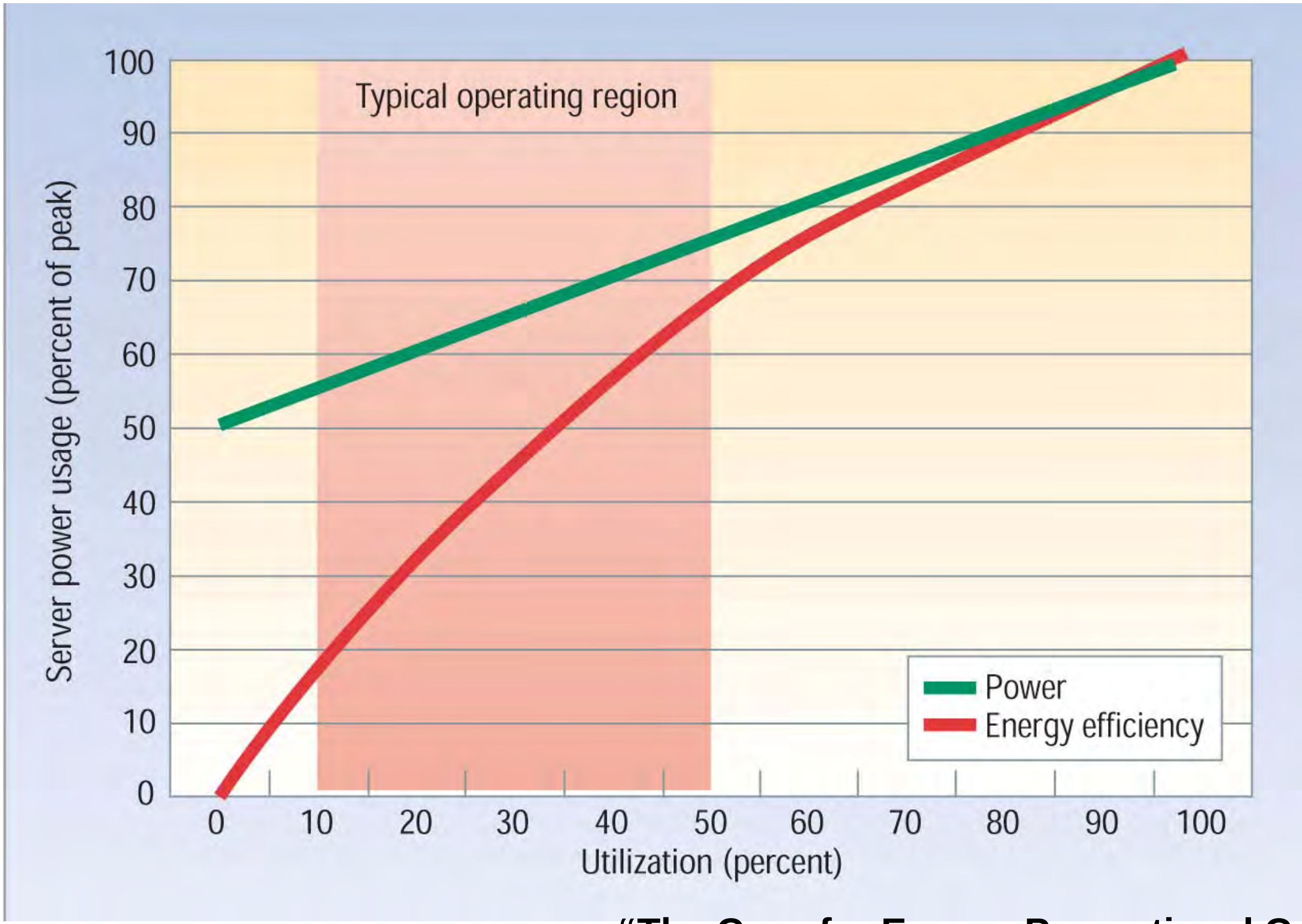
**“The Case for Energy-Proportional Computing”**



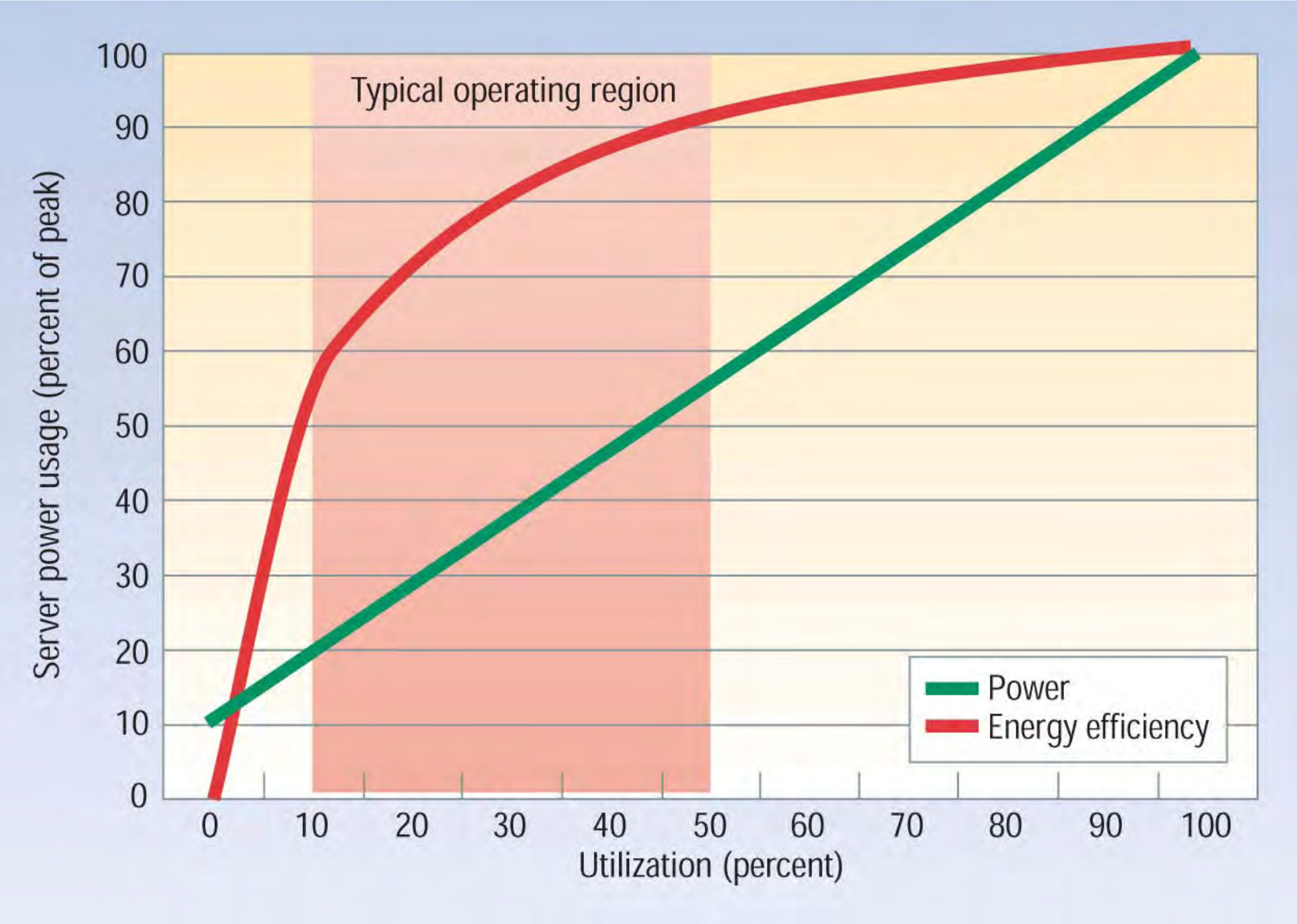
Response Time
250
200
150
100
50
0
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Utilization



**“The Case for Energy-Proportional Computing”**



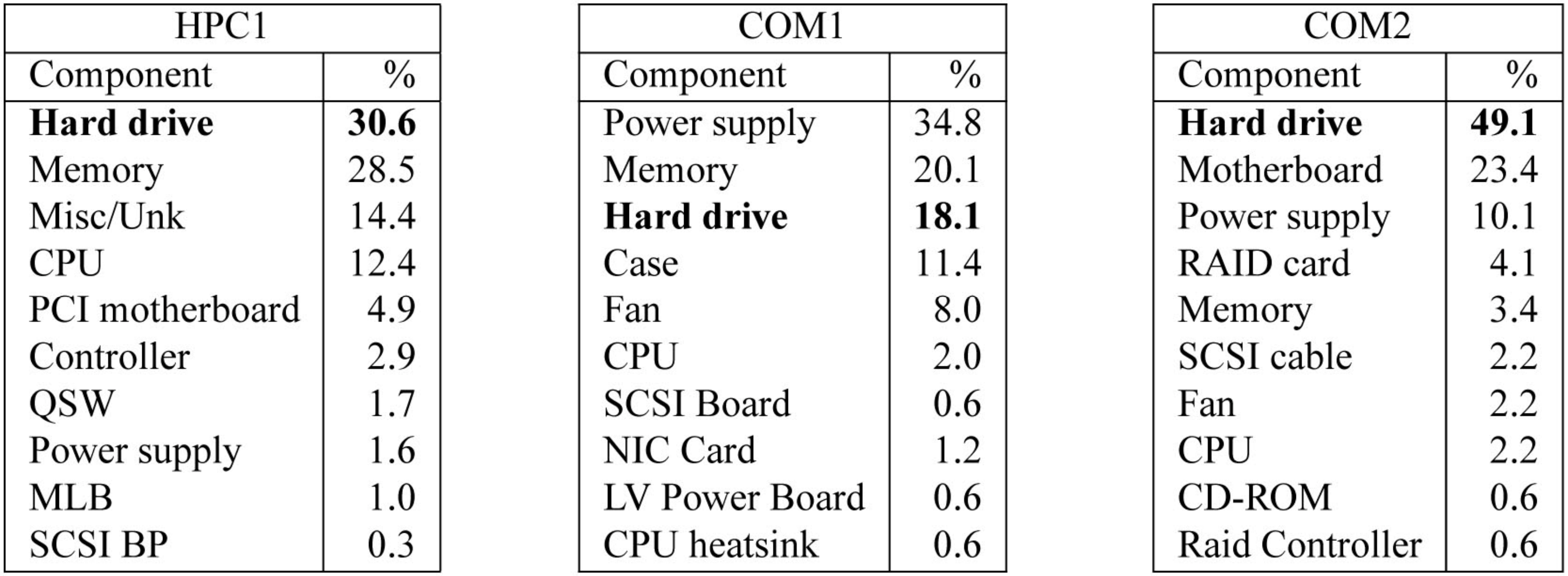
**“The Case for Energy-Proportional Computing”**



**“The Case for Energy-Proportional Computing”**

# Disk drive reliability

- Focus on disks as a commonly replaced component



| HPC1 | |
|---|---|
| Component | % |
| **Hard drive** | **30.6** |
| Memory | 28.5 |
| Misc/Unk | 14.4 |
| CPU | 12.4 |
| PCI motherboard | 4.9 |
| Controller | 2.9 |
| QSW | 1.7 |
| Power supply | 1.6 |
| MLB | 1.0 |
| SCSI BP | 0.3 |

| COM1 | |
|---|---|
| Component | % |
| Power supply | 34.8 |
| Memory | 20.1 |
| **Hard drive** | **18.1** |
| Case | 11.4 |
| Fan | 8.0 |
| CPU | 2.0 |
| SCSI Board | 0.6 |
| NIC Card | 1.2 |
| LV Power Board | 0.6 |
| CPU heatsink | 0.6 |

| COM2 | |
|---|---|
| Component | % |
| **Hard drive** | **49.1** |
| Motherboard | 23.4 |
| Power supply | 10.1 |
| RAID card | 4.1 |
| Memory | 3.4 |
| SCSI cable | 2.2 |
| Fan | 2.2 |
| CPU | 2.2 |
| CD-ROM | 0.6 |
| Raid Controller | 0.6 |

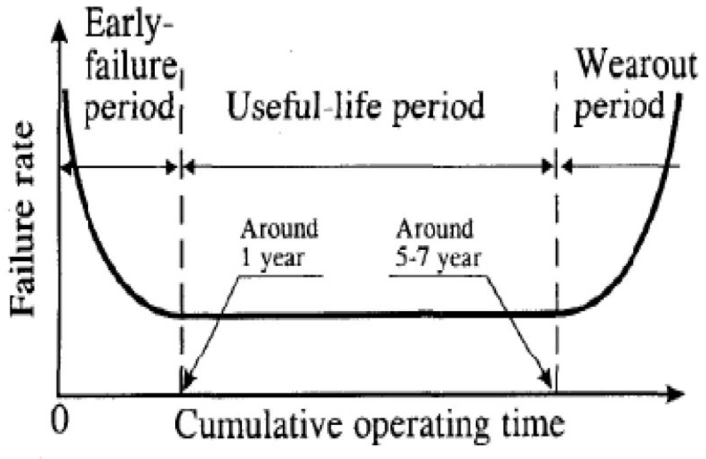
**“Disk failures in the real world”**

# Disk drive reliability

- Typical disk spec sheet MTTF is 1,000,000 hours
  - Corresponds to an annual failure rate of about 1%
- If a datacenter has 20,000 machines and each machine has 4 disks, that would be an average failure rate of more than 2 a day
- But it's worse …
  - Field replacement rates are much higher than the spec sheet MTTF would suggest
    - By a factor of 2-10 for disks less than 5 years old
    - By a factor of 30 for disks between 5 and 8 years old

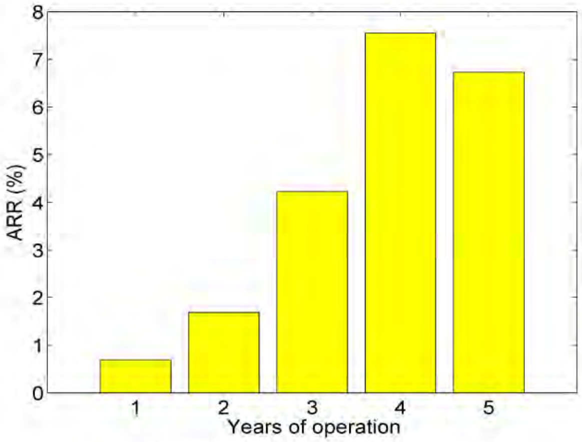
Why might this be?

- Failure rates increase annually – the "bathtub curve" doesn't represent reality

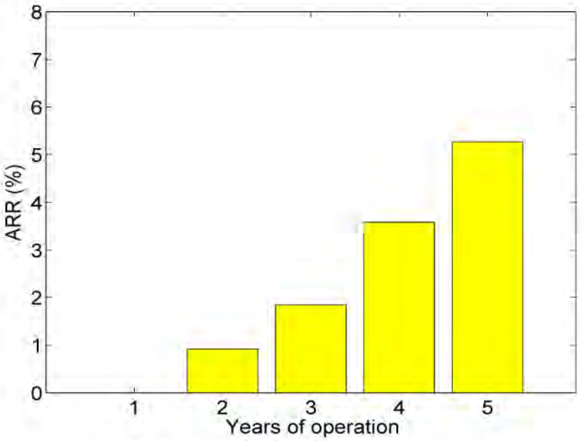


What's an example of a situation where the "bathtub curve" is realistic?

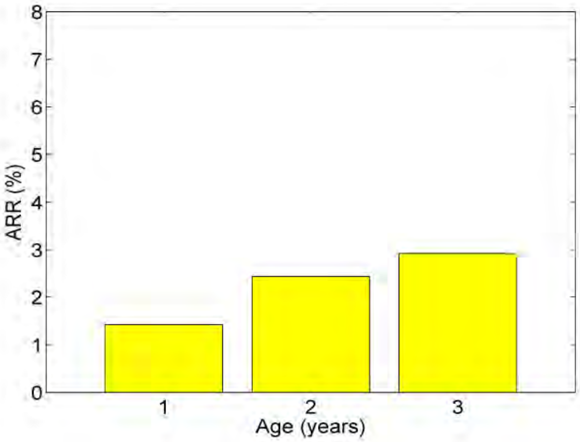
**"Disk failures in the real world"**



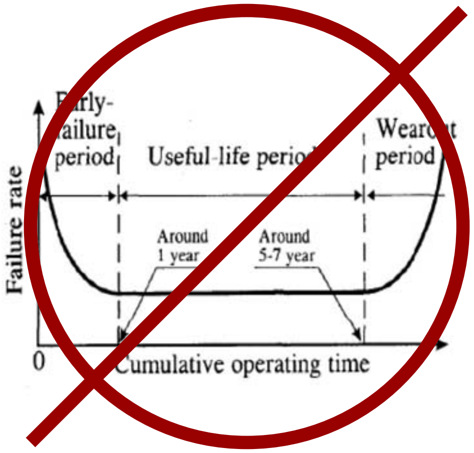
HPC1 (compute nodes)



HPC1 (filesystem nodes)



HPC4

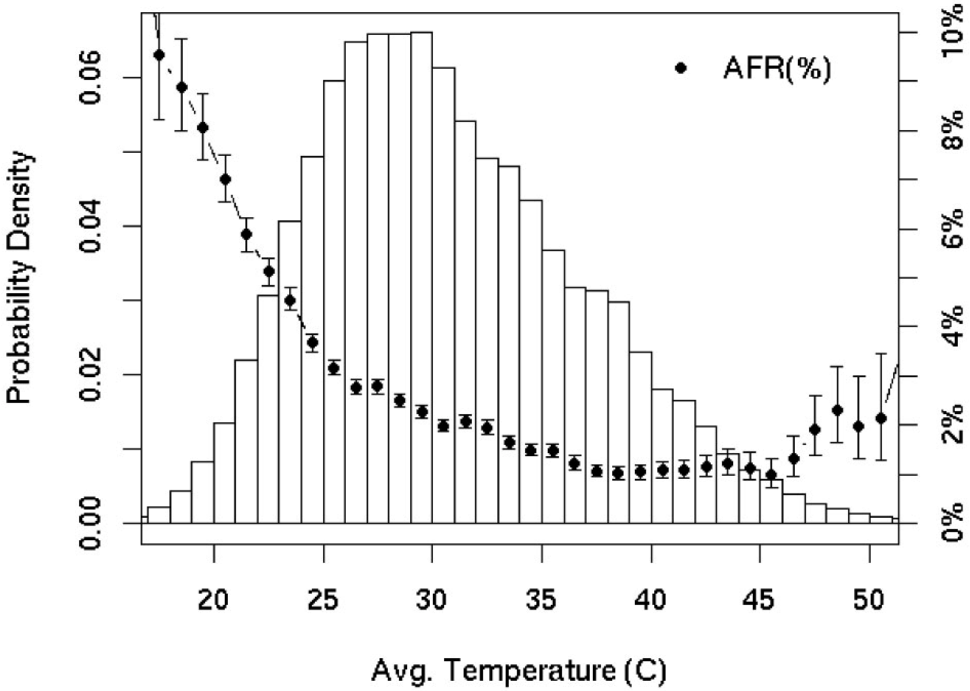


**“Disk failures in the real world”**

- Failures are clustered in time

Why might this be?

- Failures aren’t very dependent on average operating temperature



Does this contradict the previous discussion?

**"Failure Trends in a Large Disk Drive Population"**

- Failures aren't very dependent on utilization

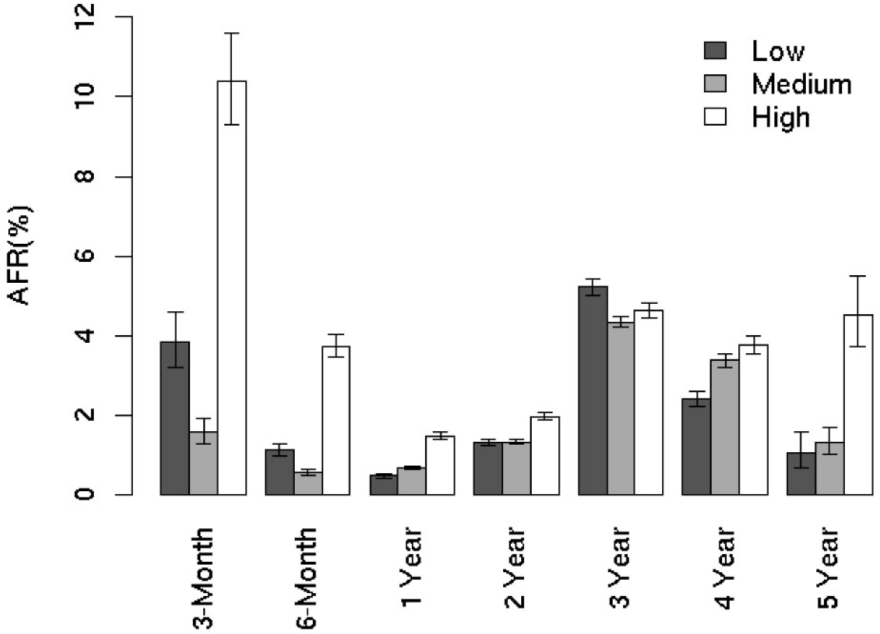
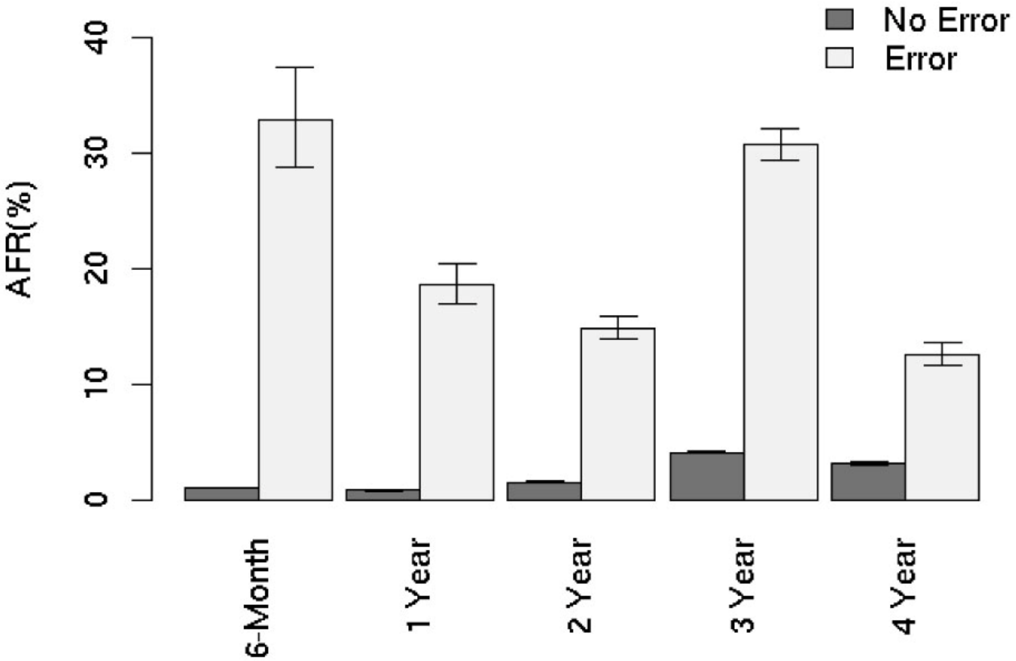


Figure 3: Utilization AFR

Except for young disks – why?

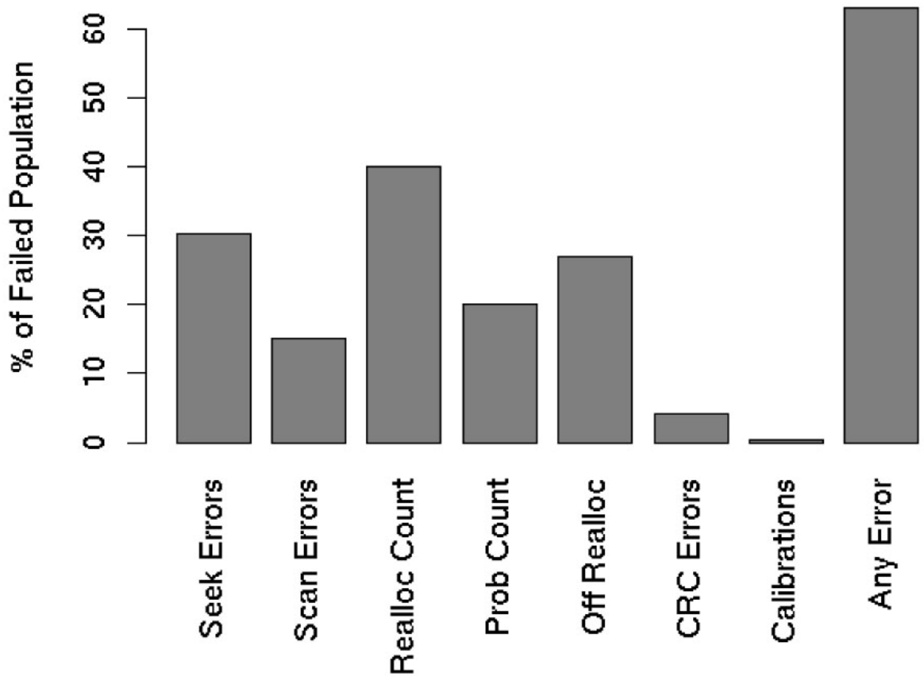
**"Failure Trends in a Large Disk Drive Population"**

- Scan errors are correlated with impending failure



**"Failure Trends in a Large Disk Drive Population"**

- But like all SMART (Self-Monitoring Analysis and Reporting Technology) parameters, scan errors don't come anywhere close to predicting all failures



**"Failure Trends in a Large Disk Drive Population"**