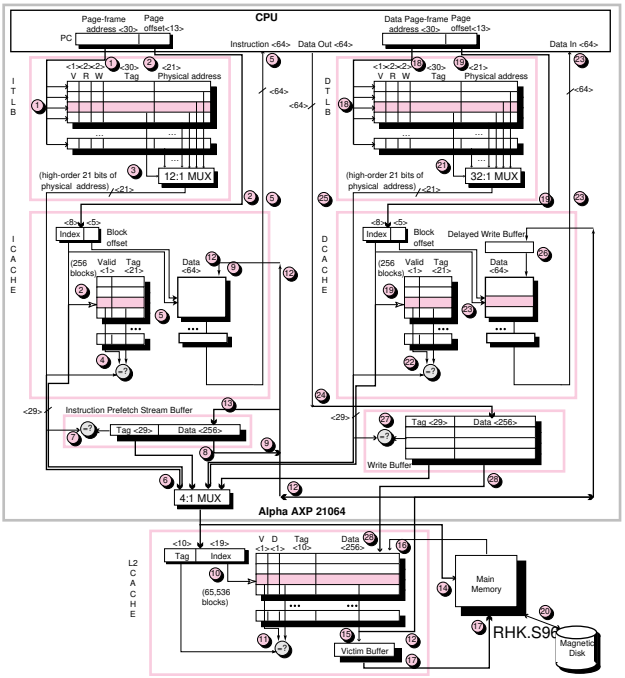
# Lecture 20: I/O— Storage Devices, Metrics, and Productivity

**Professor Randy H. Katz**
**Computer Science 252**
**Spring 1996**



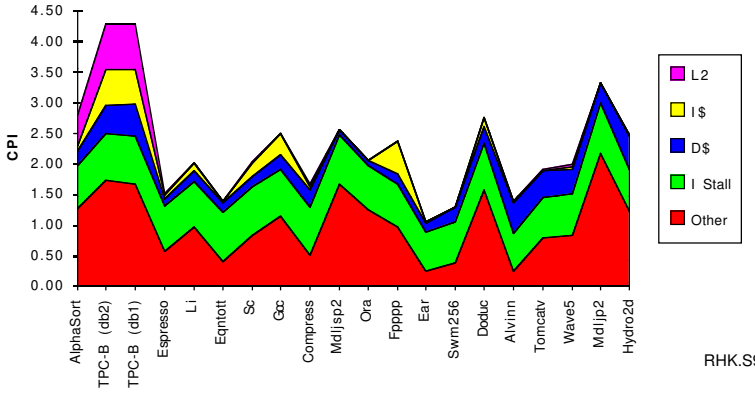
## Alpha 21064

- **Separate Instr & Data TLB & Caches**
- **TLBs fully associative**
- **Caches 8KB direct mapped**
- **Critical 8 bytes first**
- **Prefetch instr. stream buffer**
- **2 MB L2 cache, direct mapped**
- **256 bit path to main memory, 4 64-bit modules**



# Review: Alpha CPI Components

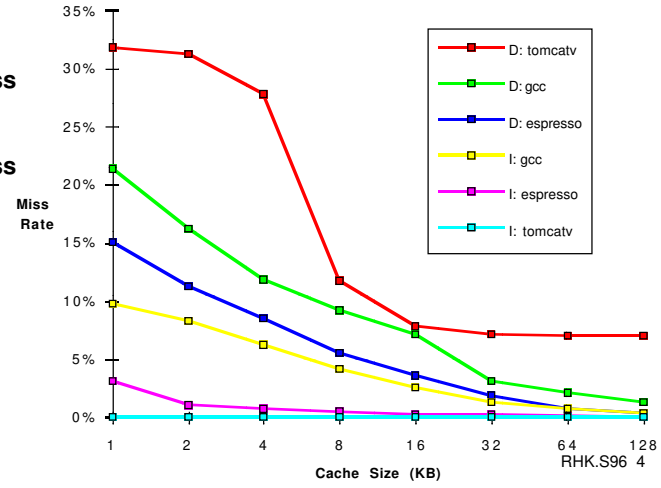
- **Instruction stalls: branch mispredict;**
- **Other: compute + reg conflicts, structural conflicts**




RHK.S96 3

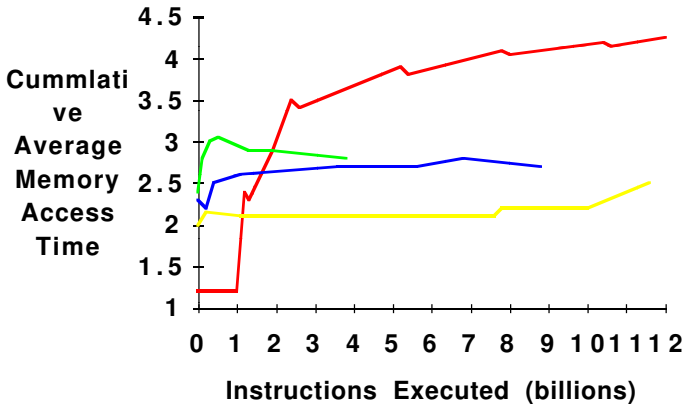
# Pitfall: Predicting Cache Performance from Different Program (ISA, compiler,...)

- **4KB Data cache miss rate 8%,12%, or 28%?**
- **1KB Instr cache miss rate 0%,3%, or 10%?**
- **Alpha vs. MIPS for 8KB Data: 17% vs. 10%**



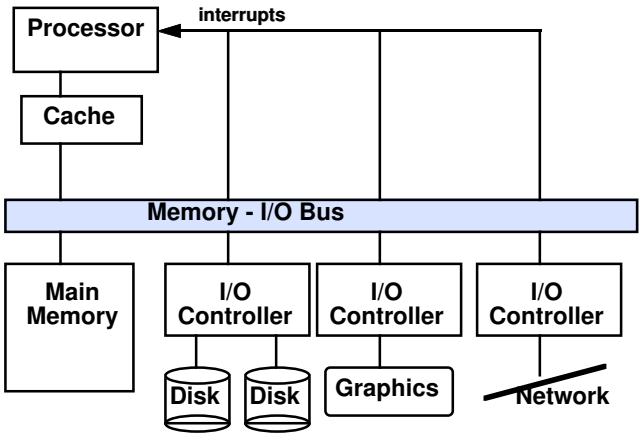

RHK.S96 4

# Pitfall: Simulating Too Small an Address Trace

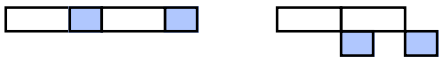



RHK.S96 5

# I/O Systems




Time(workload) = Time(CPU) + Time(I/O) - Time(Overlap)



RHK.S96 6

## Storage System Issues

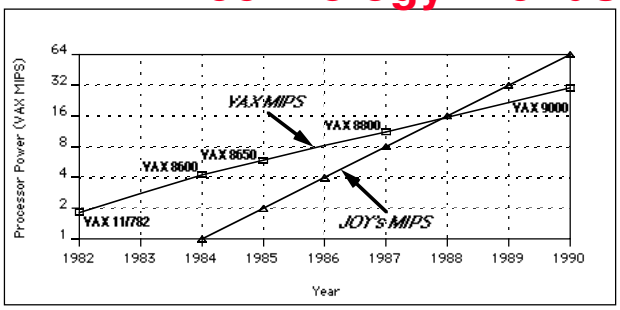
- Historical Context of Storage I/O
- Secondary and Tertiary Storage Devices
- Storage I/O Performance Measures
- A Little Queuing Theory
- Processor Interface Issues
- I/O Buses
- Redundant Arrarys of Inexpensive Disks (RAID)
- ABCs of UNIX File Systems
- I/O Benchmarks
- Comparing UNIX File System Performance



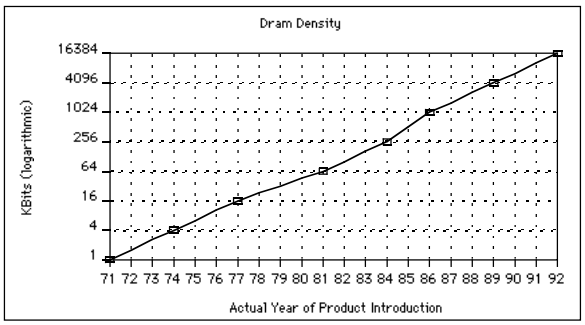
## Motivation: Who Cares About I/O?

- CPU Performance: 50% to 100% per year
- Multiprocessor supercomputers 150% per year
- I/O system performance limited by *mechanical* delays
  < 5% per year (IO per sec or MB per sec)
- Amdahl's Law: system speed-up limited by the slowest part!
  10% IO & 10x CPU => 5x Performance (lose 50%)
  10% IO & 100x CPU => 10x Performance (lose 90%)
- I/O bottleneck:
  Diminishing fraction of time in CPU
  Diminishing value of faster CPUs

# Technology Trends



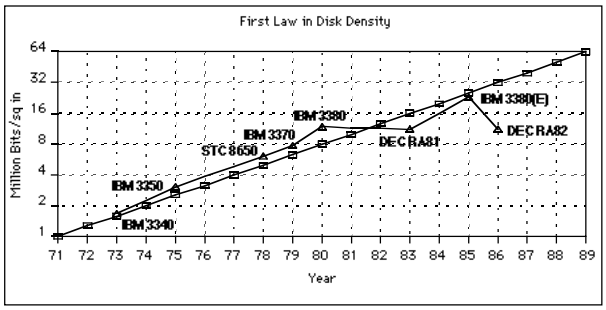

***CPU Performance***
- **Mini: 40% increase per year**
- **RISC: 100% increase per year**




***DRAM Capacity*** **doubles every 2-3 years**

RHK.S96 9

# Technology Trends




***Disk Capacity*** **doubles every 3 years**

- **Today: Processing Power Doubles Every 18 months**
- **Today: Memory Size Doubles Every 18 months(?)**
- **Today: Disk Capacity Doubles Every 18 months**
- ***Disk Positioning Rate (Seek + Rotate) Doubles Every Ten Years!***

**The I/O GAP**

RHK.S96 10

# Storage Technology Drivers

- Driven by the prevailing computing paradigm
  - 1950s: migration from batch to on-line processing
  - 1990s: migration to ubiquitous computing
    - » computers in phones, books, cars, video cameras, …
    - » nationwide fiber optical network with wireless tails
- Effects on storage industry:
  - Embedded storage
    - » smaller, cheaper, more reliable, lower power
  - Data utilities
    - » high capacity, hierarchically managed storage



# Historical Perspectives

- 1956 IBM Ramac — early 1970s Winchester
  - Developed for mainframe computers
    - » proprietary interfaces
  - Steady shrink in formfactor: 27 in. to 14 in.
    - » driven by performance demands

      higher rotation rate

      more actuators in the machine room

# Historical Perspective

- 1970s developments
  - 5.25 inch floppy disk formfactor
    - » download microcode into mainframe
  - semiconductor memory and microprocessors
  - early emergence of industry standard disk interfaces
    - » ST506, SASI, SMD, ESDI



# Historical Perspective

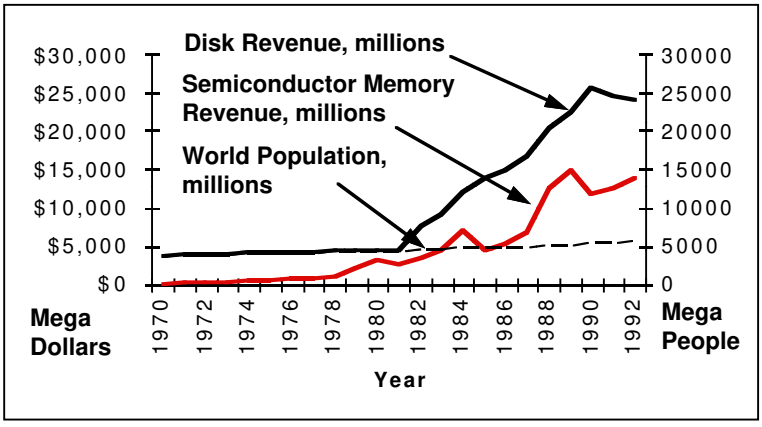
- Early 1980s
  - PCs and first generation workstations
- Mid 1980s
  - Client/server computing
  - Centralized storage on file server
    - » accelerates disk downsizing
    - » 8 inch to 5.25 inch
  - Mass market disk drives become a reality
    - » industry standards: SCSI, IPI, IDE
    - » 5.25 inch drives for standalone PCs
    - » End of proprietary disk interfaces

# Historical Perspective

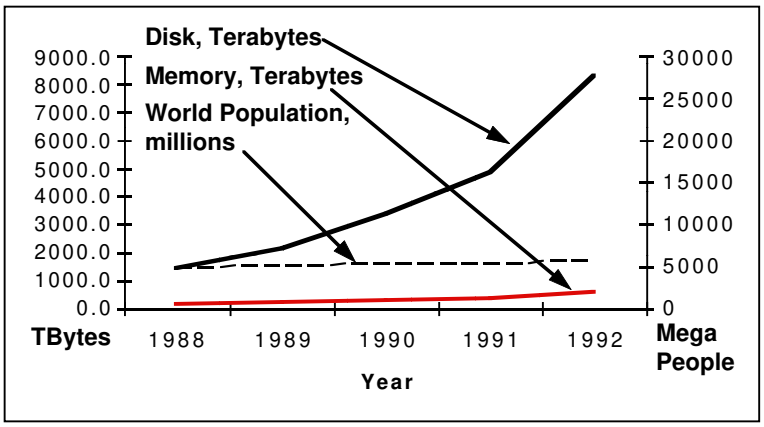
- **Late 1980s/Early 1990s:**
  - Laptops, notebooks, palmtops
  - 3.5 inch, 2.5 inch, 1.8 inch, 1.3 inch formfactors
  - Formfactor plus capacity drives market, not performance
  - Challenged by RAM, flash RAM in PCMCIA cards
    - » still expensive, Intel promises but doesn't deliver
    - » unattractive MBytes per cubic inch
  - Optical disk fails on performace (e.g., NEXT) but finds niche (CD ROM)



# Historical Perspective



# Historical Perspectives




**1.5 MBytes Disk per person on the earth sold in 1992**
**0.1 MBytes Memory per person on the earth sold in 1992** RHK.S96 17

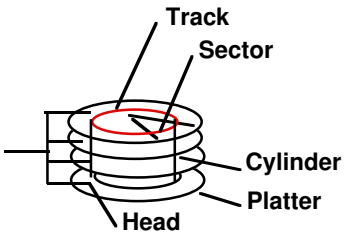
# Alternative Data Storage Technologies

| Technology | Cap (MB) | BPI | TPI | BPI*TPI (Million) | Data Xfer (KByte/s) | Access Time |
|---|---|---|---|---|---|---|
| **Conventional Tape:** | | | | | | |
| Cartridge (.25") | 150 | 12000 | 104 | 1.2 | 92 | minutes |
| IBM 3490 (.5") | 800 | 22860 | 38 | 0.9 | 3000 | seconds |
| **Helical Scan Tape:** | | | | | | |
| Video (8mm) | 4600 | 43200 | 1638 | 71 | 492 | 45 secs |
| DAT (4mm) | 1300 | 61000 | 1870 | 114 | 183 | 20 secs |
| D-3 (1/2") | 20,000 | | | | | 15 secs? |
| **Magnetic & Optical Disk:** | | | | | | |
| Hard Disk (5.25") | 1200 | 33528 | 1880 | 63 | 3000 | 18 ms |
| IBM 3390 (10.5") | 3800 | 27940 | 2235 | 62 | 4250 | 20 ms |
| Sony MO (5.25") | 640 | 24130 | 18796 | 454 | 88 | 100 ms |

RHK.S96 18

# Devices: Magnetic Disks

- **Purpose:**
  - Long-term, nonvolatile storage
  - Large, inexpensive, slow level in the storage hierarchy
- **Characteristics:**
  - Seek Time (~20 ms avg, 1M cyc at 50MHz)
    - » positional latency
    - » rotational latency
- **Transfer rate**
  - About a sector per ms (1-10 MB/s)
  - Blocks
- **Capacity**
  - Gigabytes
  - Quadruples every 3 years (aerodynamics)



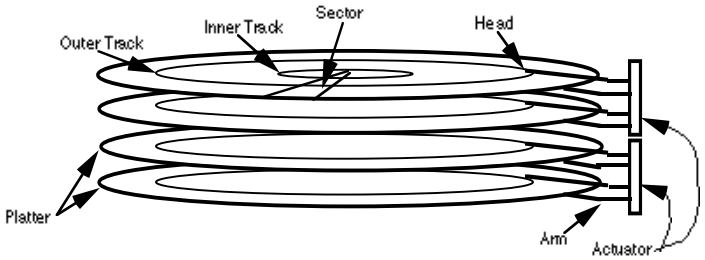

3600 RPM = 60 RPS => 16 ms per rev
ave rot. latency = 8 ms
32 sectors per track => 0.5 ms per sector
1 KB per sector => 2 MB / s
32 KB per track
20 tracks per cyl => 640 KB per cyl
2000 cyl => 1.2 GB

Response time
= Queue + Controller + Seek + Rot + Xfer

Service time



# Disk Device Terminology




**Disk Latency = Queuing Time + Seek Time + Rotation Time + Xfer Time**

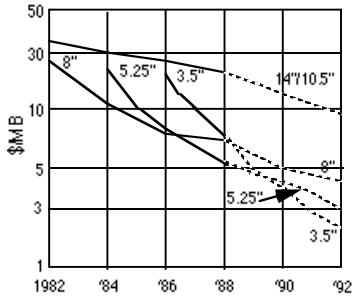
***Order of magnitude times for 4K byte transfers:***

**Seek: 15 ms or less**

**Rotate: 8.3 ms @ 3600 rpm (4.2 ms @ 7200 rpm)**

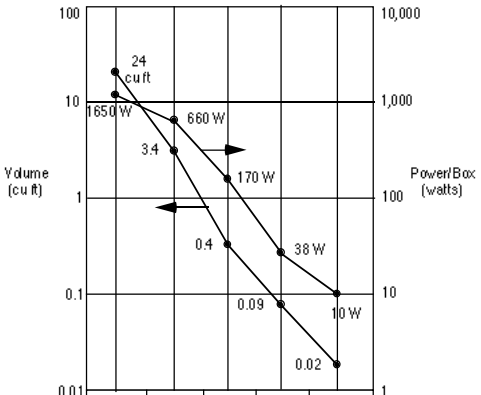
**Xfer: 2 ms @ 3600 rpm (1 ms @ 7200 rpm)**

# Advantages of Small Formfactor Disk Drives




**Low cost/MB**
**High MB/volume**
**High MB/watt**
**Low cost/Actuator**

***Cost and Environmental Efficiencies***




| | 14" | 10.5" | 8" | 5.25" | 3.5" |
|---|---|---|---|---|---|
| MB | 7,500 F | 1,320 F | 1,000 | 765 | 200 |
| MB/watt | 4.54 | 2.0 | 5.88 | 20.13 | 20.0 |
| MB/cu ft | 313 | 388 | 2,500 | 8,500 | 10,000 |

RHK.S96 21

# Tape vs. Disk

- **Longitudinal tape uses same technology as hard disk; tracks its density improvements**

- **Inherent cost-performance based on geometries: fixed rotating platters with gaps**
  **(random access, limited area, 1 media / reader)**

**vs.**

**removable long strips wound on spool**
**(sequential access, "unlimited" length, multiple / reader)**

- **New technology trend:**
  **Helical Scan (VCR, Camcoder, DAT)**
  **Spins head at angle to tape to improve density**

RHK.S96 22

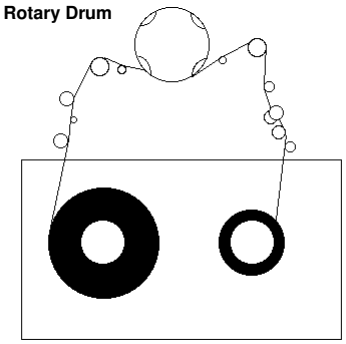
# Example: R-DAT Technology

**Rotating (vs. Stationary) head Digital Audio Tape**

- **Highest areal recording density commercially available**
- **High density due to:**
  - **high coercivity metal tape**
  - **helical scan recording method**
  - **narrow, gapless (overlapping) recording tracks**
- **10X improvement capacity & xfer rate by 1999**
  - **faster tape and drum speeds**
  - **greater track overlap**

RHK.S96 23

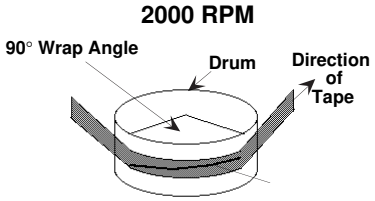
# R-DAT Technology




**Four Head Recording**

**Tracks Recorded ±20° w/o guard band**

**Read After Write Verify**

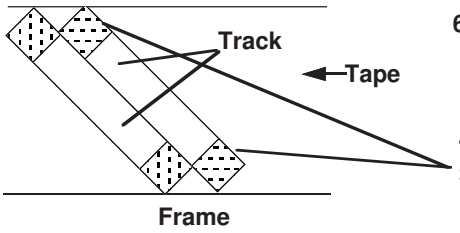



**Helical Recording Scheme**

RHK.S96 24

# R-DAT Technology

DDS ANSI Standard (HP, SONY)




65% of Track is Data Area
70% Data Bytes
30% Bytes Parity Plus
Reed-Solomon Codes

Track Finding Area (Servo)
Subcode Area (Index)
Margin Area

Block
Track (2900 Data Bytes)
Frame (2 Tracks)
Group (22 Frames + Optional Group ECC, 128K bytes)

Theoretical Bit Error Rates:

- w/o group ECC: one in $10^{26}$
- w/ group ECC: one in $10^{33}$

RHK.S96 25

# Optical Disk vs. Tape

| | Optical Disk | Helical Scan Tape |
|---|---|---|
| Type | 5.25" | 8mm |
| Capacity | 0.75 GB | 5 GB |
| Media Cost | $90 - $175 | $8 |
| Drive Cost | $3,000 | $3,000 |
| Access | Write Once | Read/Write |
| Robot Time | 10 - 20 s | 10 - 20 s |

Media cost ratio optical disk vs. helical tape
= 75 : 1 to 150 : 1

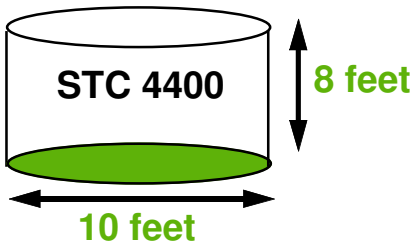
RHK.S96 26

# Current Drawbacks to Tape

- **Tape wear out:**
  - **Helical 100s of passes to 1000s for longitudinal**
- **Head wear out:**
  - **2000 hours for helical**
- **Both must be accounted for in economic / reliability model**
- **Long rewind, eject, load, spin-up times; not inherent, just no need in marketplace (so far)**



# Automated Cartridge System




**6000 x 0.8 GB 3490 tapes = 5 TBytes in 1992**
**$500,000 O.E.M. Price**

**6000 x 20 GB D3 tapes = 120 TBytes in 1994**
**1 Petabyte (1024 TBytes) in 2000**

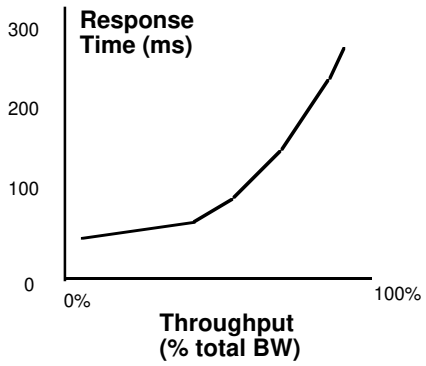
# Relative Cost of Storage Technology—Late 1995

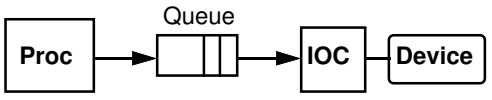
| | | | |
|---|---|---|---|
| **Magnetic Disks** | | | |
| 5.25” | 9.1 GB | $2129 | $0.23/MB |
| 3.5” | 4.3 GB | $1199 | $0.27/MB |
| 2.5” | 514 MB | $299 | $0.58/MB |
| **Optical Disks** | | | |
| 5.25” | 4.6 GB | $1695+199 | $0.41/MB |
| **PCMCIA Cards** | | | |
| Static RAM | 4.0 MB | $700 | $175/MB |
| Flash RAM | 40.0 MB | $1300 | $32/MB |
| | 175 MB | $3600 | $20.50/MB |

RHK.S96 29

# Disk I/O Performance

**Metrics:**
- **Response Time**
- **Throughput**




**Response time = Queue + Device Service time**

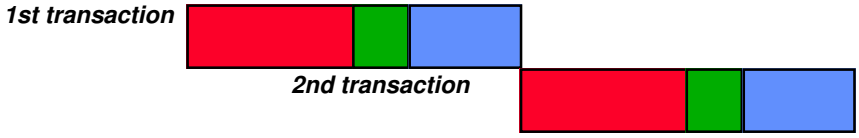
RHK.S96 30

# Response Time vs. Productivity

- **Interactive environments:**

  **Each interaction or *transaction* has 3 parts:**

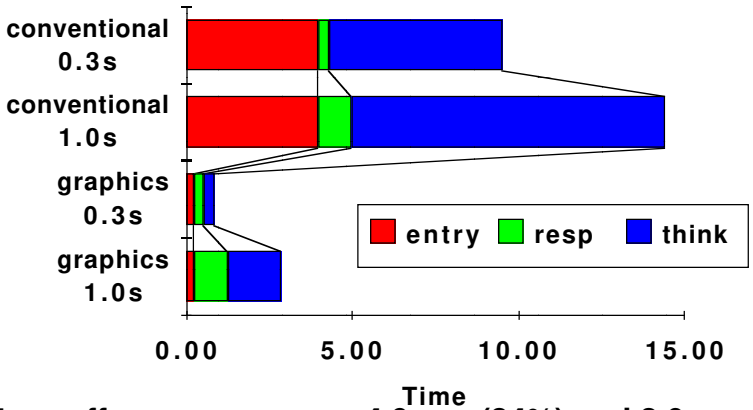
  - ***Entry Time*: time for user to enter command**
  - ***System Response Time*: time between user entry & system replies**
  - ***Think Time*: Time from response until user begins next command**




- **What happens to transaction time as shrink system response time from 1.0 sec to 0.3 sec?**
  - **With Keyboard: 4.0 sec entry, 9.4 sec think time**
  - **With Graphics: 0.25 sec entry, 1.6 sec think time**



# Response Time & Productivity




- **0.7sec off response saves 4.9 sec (34%) and 2.0 sec (70%) total time per transaction => greater productivity**
- **Another study: everyone gets more done with faster response, but novice with fast response = expert with slow**