Storage Technologies Overview

Low-Power Computing
Carnegie Mellon University
David Andersen

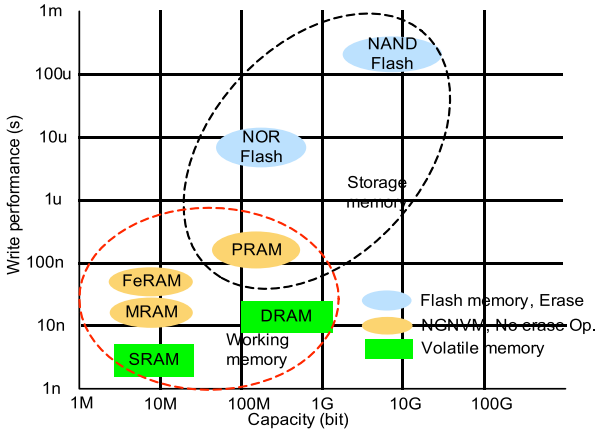
What ?s should we ask?

- Volatile vs. Non-Volatile (pull the plug)
- Density: bits / area
- Cost: bits / $
- Durability / lifetime
  - How long does data last? How many write cycles?
- Speed - random, sequential, small, large, reads, writes

How can we store things?

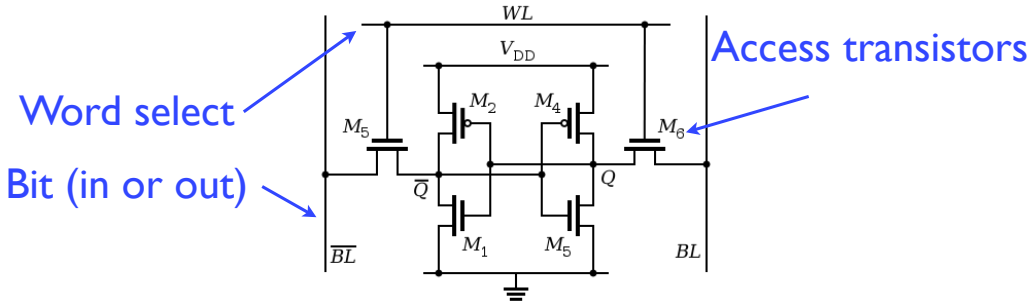
- Charge
- Magnetic polarization (electron spin)
- Light in a loop
- Physical state / chemical bonds / etc.
  - punch cards, CD-ROMs, DVDs,
  - holographic storage, DNA..

Disk



- Source: Samsung Electronics
  "A PRAM and NAND Flash Hybrid Architecture for High-Performance Embedded Storage Subsystems"

Transistors: SRAM



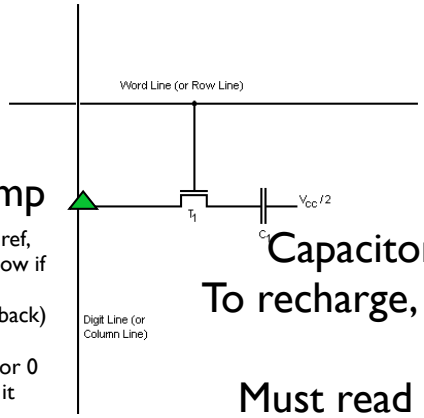
- Typical SRAM uses 6 transistors to store one bit -- 4 store the data, 2 to control read or write. The 4 are basically a flip-flop.
- Read: Transistors drive BL high or low actively (Fast!). Write: complicated. :) (set BL low, set to bit, toggle WL)
- Idle: Inverters reinforce each other; nothing changes state - low power
- 

Tiny image from HowStuffWorks

SRAM

- High power when used @ high bandwidth
- Density: Bad. 6 transistors per bit. Ow!
- They're made of transistors -- easy to put directly on-chip (nice!) -> CPU caches
- Idle power draw is the same as CPUs - leakage current. High-k can reduce this... for now. :)

Capacitors: DRAM



sense amp

Compares line to ref, drives it to high or low if > or < the ref. (Think positive feedback)

Result 1: Strong 1 or 0 signal - can read it

Result 2: Refreshes value in capacitor (drives it all the way to 1 or 0)

Capacitors leak.
To recharge, must read.

Must read *every cell*

base figure from Thomas Schwarz, augmented.

DRAM properties

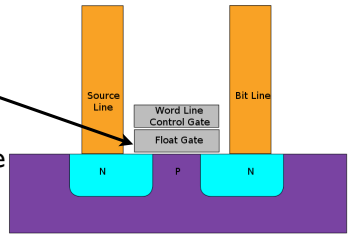
- Pretty dense. 1C + 1T per bit.
- Pretty fast - ~20ns random access, but very high bandwidth due to very parallel internal structure (1.6 * 10^9 32 bit words per second). Reads & writes both fast.
- Cost - about $10/GB
- Power: High. Up to 5-10W per DIMM. High (100s of mW) power draw when idle -- refresh bandwidth
- Leakage increases as size decreases - so refresh happens more frequently (idle power goes up relative to active)

Example DRAM

- Micron 8GB FBDIMM, 240pin DDR2, 4 Gb/sec transfer rates (serial!)
- needs +1.8v, +3V for onboard controller
- 4.1W in lowest idle state
- 11.5W active
  - This is fast, ECC, server ram - NOT low power, but nice example..
  - laptop 4GB SO-DIMM DDR3 @ 1333Mhz: ~1.5W active, 200mW power-down, 4.5W reading from all banks interleaved

Transistors again: NAND Flash

Charge stored here

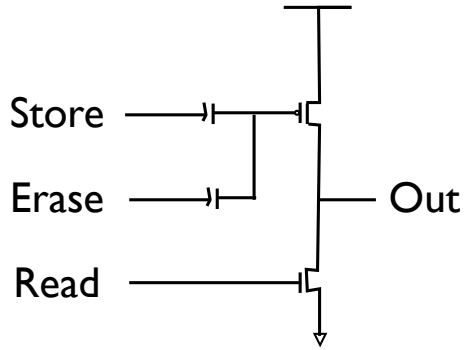


To charge: Apply higher-than-normal control voltage

To reset: Apply even higher source voltage or negative control V (~10-20V)

- Obligatory stolen-from-Wikipedia image
- Flash stores one (or a few) bits per transistor -- "floating gate"
- So really, it's transistors-as-capacitors
- Key thing: High electric field causes electron tunneling (to or from float gate)
  - And causes the insulator oxide to break down a little bit...
  - (Caveat: Newer flash uses hot carriers, not tunneling - faster, but hurts reliability a bit)

Schematic view of flash



SLC or MLC?

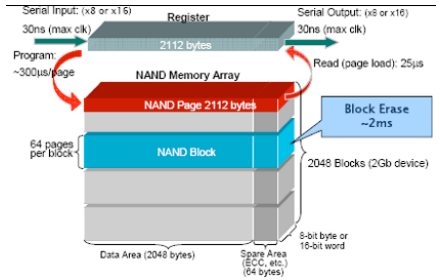
- Low power to read, medium power to write, higher power to erase. (Remember sensor paper - higher power to write; most writes require erases)
- Can store one bit (SLC) or multiple bits (MLC) per transistor by storing multiple levels of charge
- MLC - higher density, more careful programming. Smaller amount of charge difference causes bit flips -> less robust.

	SLC	MLC
Density	16Mbit	64Mbit
Read	100ns	150ns
Block	64K	128K
Endurance	100k cycles	10k cycles
Temp range	larger	smaller

Table from SuperTalent Technology whitepaper, "SLC vs. MLC"

Tradeoffs in flash

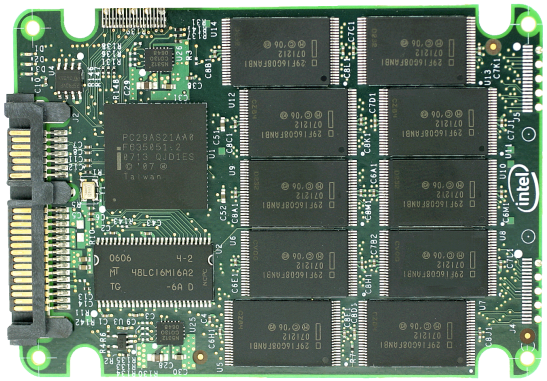
- Cost/Density vs. reliability
- Power vs. Speed - faster to program with higher voltage, draws more power.
- Speed vs. lifetime: thicker gate oxide retains data longer, slower to program



- Yoinked from comms daily
- Erase: 2ms

NAND Flash SSD is a little computer

- Storage: flash chips
- Access: multiple independent access channels
- Interface: SATA
- Controller: computer + RAM
  - Processes cmds
  - Drives channels
  - Write behind
  - Allocation
  - Wear leveling



Flash summary

- $1-ish per gigabyte
- No power when idle. 1W SSD = 100MB/s
- With wear leveling, lifetime of ~3Y continuous writing with good flash
- Data degrades - it's a capacitor! Maybe a year? Reads cause somewhat faster degrade.
- Fast sequential writes, slow random writes (erase block), V fast sequential reads, quite fast random reads (but first byte more expensive)

Magnets: MRAM

- Store a bit as magnetic polarization -
- ---current--> Magnet |barrier| Magnet -->
- Resistance changes if polarities aligned or opposite
  - (Cute point: This is used in hard drives, too)
- The hype: Density and speed of DRAM, but non-volatile and no refresh current. Latency almost as fast as SRAM.
- The reality: A few MRAMs now exist. 180nm process. Many challenges in shrinking it. *Far* less dense than flash (coming later...)

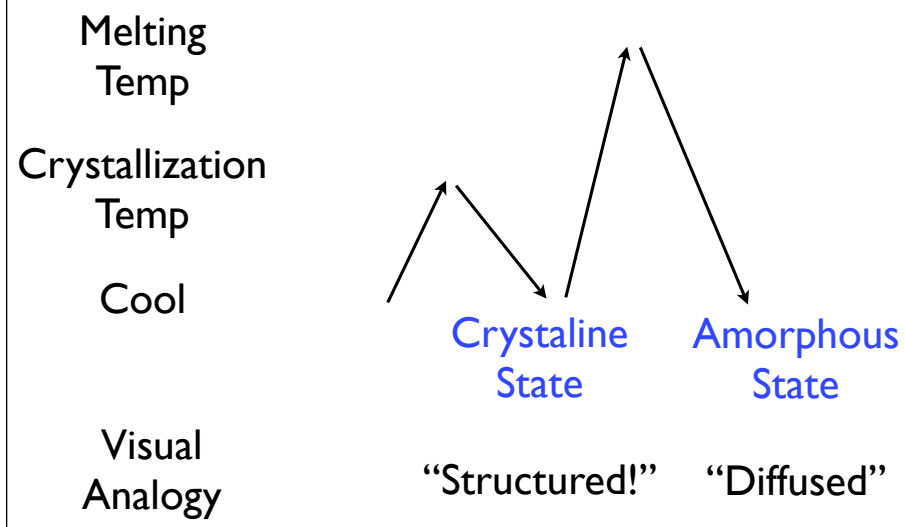
Everspin MRAM

- (spin-off, no pun intended, from FreeScale. FreeScale was a spin-off from Motorolla)
- 4 Mbit MRAM chip (MEGA-bit; Micron will sell you a 4 GIGA-bit DRAM chip)
- 35ns read/write cycle (~roughly DRAM)
- Unlimited read/write cycles
- TPD of 0.6W for 4zMbit; read current ~80mA, write ~165mA, standby 12mA (but you can turn it off completely and it keeps its data, like flash)
- Cost... ?

Physical State: Phase-Change Memory

- Idea: Find a material that changes phase with (heat, voltage, magnetic field, etc).
  - eg, amorphous vs. crystaline structure
  - (Analogy: gas vs. liquid, but that takes lots of power)
- One such material: chalcogenide glass.
  - Huh? -- It's the same stuff in rewriteabe DVDs.
  - Add sulfur, selenium, etc., to glass -> GeSbTe (Germanium-Antimony-Tellurium) (Germanium == same group as Silicon; GeO2 is a glass just like SiO2. It's popular for doping fiber optic cables.)

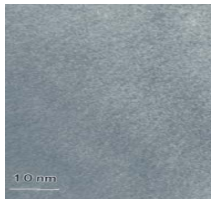
PCM diagram



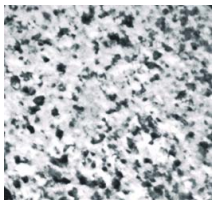
Different electrical and optical characteristics

Amorphous Phase

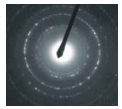
Crystalline Phase



TEM Images



Electron Diffraction Patterns



Material Characteristics

- Short-range atomic order
- Low free electron density
- High activation energy
- High resistivity

- Long-range atomic order
- High free electron density
- Low activation energy
- Low resistivity

Credit: Ovonyx corp.

Making PCM

- Have to be able to rapidly heat & cool material
  - Luckily, that gets *easier* with scaling down (from a power perspective)
  - Maybe 100ns to cool down. Very small area. :)
- To heat: big current pulse (1mA)
  - (Requires constant current source)

PCM Characteristics

- Read: *Fast*. Measuring resistance (or optically). Speed will depend a lot on other decisions - MLC vs SLC, serial vs. parallel structure, etc. But nothing fundamentally slow.
- Non-volatile
- Writing: Probably slower than DRAM, but doesn't have to erase like Flash
- Density is tougher than DRAM
- Lifetime - 10x that of flash?
- The reality: 1 shipping PRAM part;
  many companies say they have samples at surprisingly nice density (512Mbit, etc.)
- Similarish comments apply about FeRAM (uses

Separation of Read & Media

- Previous technologies directly read data out of cells
  - ~1 transistor per cell, all were lithographic processes
- Why does the reading technology have to be attached? Many to 1.... as long as you don't mind physically moving the read device.
- Density & cost vs. moving parts - speed & power. (speed particularly == latency)

2006 Laptop Drive

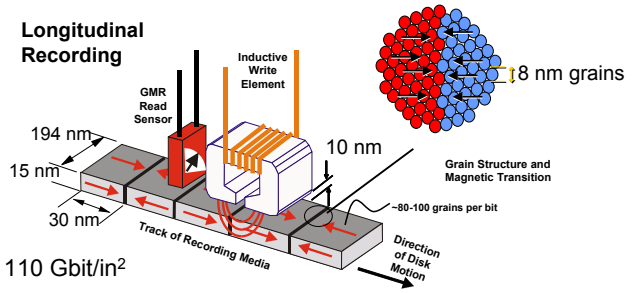
Perpendicular Recording Notebook Drive

Drive Capacity (GB)	160
Number of Discs	2
Capacity (GB/disc)	80
KTPI (avg)	147
KBPI (nom)	885
Product Areal Density (Gb/in^2)	130.1
Transfer Rate (MB/sec)	44
RPM	5400
Seek Time (ms)	
Average Read	10
Average Write	11

slide credit: Mark Kryder, Seagate

Magnetic Disk

Recording Basics

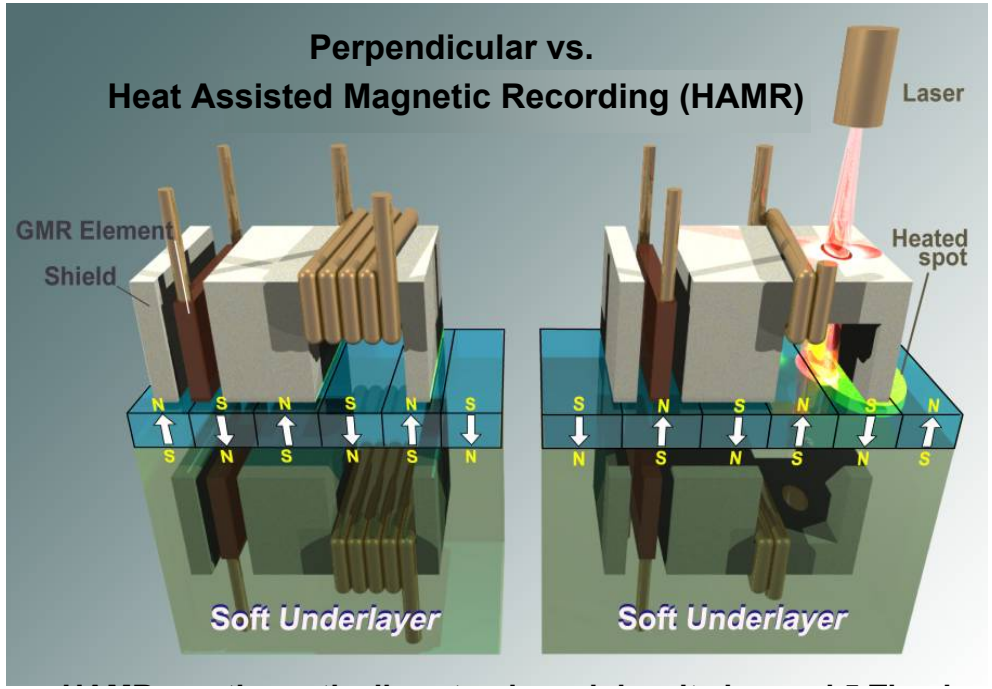


110 Gbit/in^2

(Gigabits per square inch)

- **Bit dimensions 30 x 194 x 15 nanometers**
- **Head-media spacing of 10 nanometers**

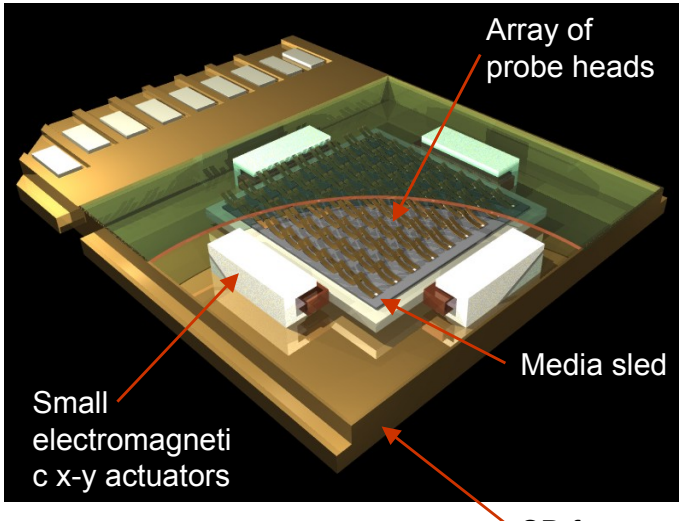
Image credit: Mark Kryder, Seagate


Perpendicular vs.
Heat Assisted Magnetic Recording (HAMR)
Laser
GMR Element
Shield
Heated spot
N S N S N S
S N S N S N
Soft Underlayer
S N S N S N
N S N S N S
Soft Underlayer

Disk Power

- Spin down: ~4 seconds
- Spin up: ~5 seconds
- Seagate Cheetah 15.5 (15k RPM), 300GB
  - 2.0ms avg latency (impressive)
  - ~100MB/sec transfer - 4 discs, 8 heads
  - 17.6W active, 12.5W idle
- Seagate Momentus - laptop, 5400 RPM, 500GB
  - 1.54W seek, 2.6W read, 2.85W write
  - 0.81 W idle (spinning)
  - 0.22W standby (spun down, electronics up)
- Power propto RPM squared, and 4th power of radius (drag)
  - Capacity propto radius ^2... doh -> 2.5" disks.

MEMS



slide credit: Mark Kryder, Seagate

MEMS Discussion

- Lots of read heads - massively parallel?
- Smaller movement needed - lower latency, lower power
- Doesn't move when idle - low-power
- Uses normal-ish magnetic media
- Can't buy one off the shelf...

Optical Media

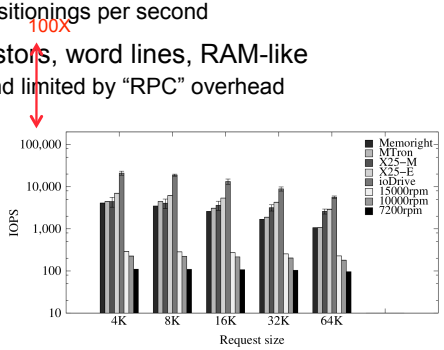
- We saw it before - same tech as PCM for -RW; organic dye for write-once
  (write-never: pressed physically. cheap!)
- But on spinning media like hard drive
- Lifetime unknown (years, but how many?)
- Speed poor - seek latency and slower rotational velocity (10k RPM limit - plastic)
- Density low: ~7GB for a DVD vs. 250GB for an equivalent drive platter, but media pretty cheap.
- Denser media - slower write times (have to focus laser on spot long enough to heat it) -- 6x for DVD-RW (=~ 3400 RPM).

The Far-Off

- Holographic storage (write-once, 3D, who knows...)
- FeRAM, Nanotube-based RAM, ...

Performance

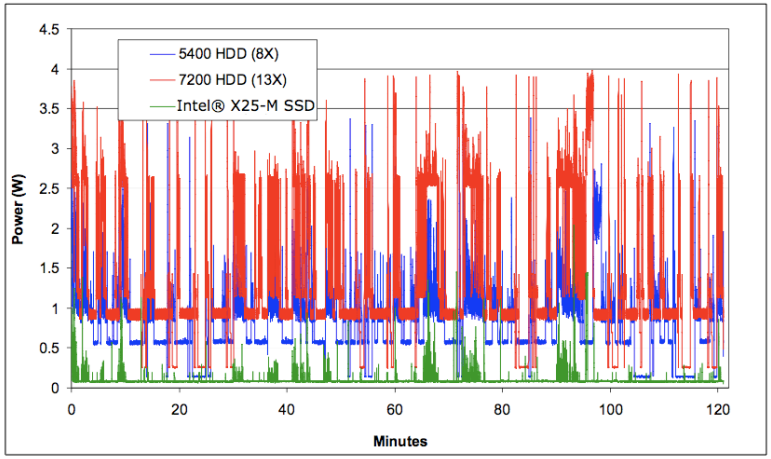
- Mechanical disk seeks are slow
  - A few milliseconds to move head, a few more for rotate
  - At most 100-200 positionings per second
- Solid-state is transistors, word lines, RAM-like
  - Accesses per second limited by “RPC” overhead
- Measurements:
  - Random read
  - SSD NAND flash vs magnetic disk
  - Laptop (Mtron, Memoright)
  - Server (Intel X25, Fusion-io iodrive)



SSDs use lower power than disks

- Nothing needs to be spun (always!)
- Nothing needs to be seeked
- Serial bits use faster logic than parallel

Intel® Mainstream SATA SSDs Save Power: SATA Power Rails With 2 Hour Mobile Workload



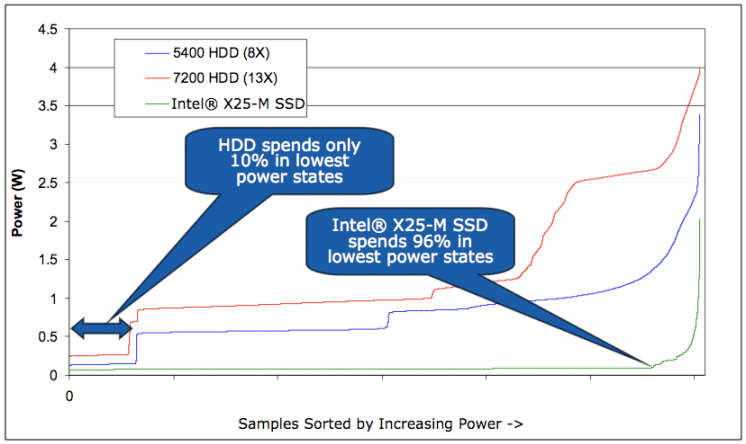
For more detail see EBLS003: "Extending Battery Life of Mobile PCs Using Intel® High Performance SATA Solid-State Drives"

Source: Intel

SSDs use lower power than disks (2)

- SSD uses almost no power most of the time
- Disks take time to spin down - have to be careful.

Intel® Mainstream SATA SSDs Save Power: SATA Power Rails With 2 Hour Mobile Workload



Source: Intel

Slightly old but representative #s

Drive Type	Model	Erase cycles	Capacity	Price	Dollars/Gigabyte	Access Time
Consumer SATA SSD	MTron Mobi	100,000	16 GB	$370	$23.13	0.1 msec
Consumer SATA SSD	Memoright GT	100,000	16 GB	$510	$31.88	0.1 msec
Enterprise SATA SSD	Intel X25-M	10,000	80 GB	$730	$9.13	0.085 msec
Enterprise SATA SSD	Intel X25-E	100,000	32 GB	$810	$25.31	0.085 msec
Enterprise PCIe SSD	FusionIO ioDrive	100,000	80 GB	$2400	$30.00	0.05 msec
7200 RPM SATA Drive	Seagate Barracuda 7200.11	∞	750 GB	$110	$0.15	4.2 msec
10K RPM SCSI 320 Drive	Seagate ST3300007LW	∞	300 GB	$350	$1.17	4.7 msec
15K RPM SCSI 160 Drive	Seagate ST3300655LC	∞	300 GB	$425	$1.42	3.9 msec