

THE ERA OF STARDOM FOR THE CUMBERSOME AND SLOW PARALLEL-STORAGE INTERFACE IS DRAWING TO A CLOSE AS NIMBLER, YOUNGER SERIAL CONTENDERS TAKE THE STAGE.



Speedy simplicity

THE ELECTRONICS PRESS frequently discusses Intel's influence on the evolving PC architecture, by virtue of its dominant microprocessor-market share. Intel's supremacy even occasionally gets the attention of regulatory

bodies in the United States and other countries. Less attention-getting but perhaps even *more* influential is Intel's prominent position in core-logic chip sets. The company's market-share leadership in this arena extends across all PC proliferations: notebooks, desktops, workstations, and servers.

Now-commonplace core-logic functions, such as integrated 3-D graphics and the AGP (accelerated graphics port) external graphics bus, DDR SDRAM, PCI internal peripheral-expansion sockets, USB 1.1 and 2.0 external peripheral-expansion ports, and AC'97 for audio, hit their strides when Intel rolled support for them into its core logic. IEEE1394, in contrast, has never achieved significant momentum in the PC architecture because it never found its way onto an Intel chip set. And, as Rambus learned, Intel-driven market success quickly evaporates once Intel subsequently withdraws that endorsement.

The mass-storage interface became the latest PC building block to undergo an Intel-initiated transformation when Intel introduced its Serial ATA (Advanced Technology Attachment)-inclusive i865 and i875 chip sets in mid-April 2003. Chip-set competitors, such as Acer

Labs, ATI Technologies, Nvidia, SiS, and Via Technologies, quickly responded with their own SATA (Serial ATA)-inclusive plans and products, which, in some cases, a PHY (physical layer)-only SATALite chip from Silicon Image accelerates.

PCI-based full-featured SATA controllers are now available from many suppliers; note that the 32-bit, 33-MHz PCI bus's 133-Mbyte/sec bandwidth holds back the full potential of first-generation SATA's 150-Mbyte/sec theoretical peak transfer rate. PCI Express-based variants now under development should relieve this constraint. SATA-to-PATA (parallel-ATA) bridge chips bring legacy mass-storage peripherals into the serial interface era, albeit with similar bottlenecks to those of their host-side bus-bridge brethren.

HISTORICAL PRECEDENTS, FUTURE TRENDS?

Serial-interface stalwart Fibre Channel isn't standing still in the face of upstart SATA's surge, and SCSI is also catching a serial wave with SAS (Serial Attached SCSI). To understand the motivations behind the industry's serial embrace and how the various interface contenders may sort out in the future, it's helpful to first take a

At a glance34
SCSI-derived transformations...34
External-storage revolutions.....36
For more information.....38

look back in time and see how things got to this point. Be careful, though, because trends beyond the interface itself may change the future rules of the game (**Reference 1**).

You might be surprised to learn that Fibre Channel wasn't the first serial-storage interface. That honor goes to the IBM-developed SSA (Serial Storage Architecture), which delivered basic 20-Mbyte/sec transfer rates that, due to the interface's dual-ported and full-duplex characteristics, translated to 80-Mbyte/sec peak speeds. Although standardization efforts began in 1992, SSA never achieved a critical mass of industry support. A speed of 80 Mbytes/sec may seem slow in comparison with today's interfaces, but keep your opinions in perspective: In 1992, SCSI-1 ran at 5 Mbytes/sec, SCSI-2 hit 20-Mbyte/sec speeds, and not-yet-standardized ATA (first-generation ATA is also known as IDE, for integrated drive electronics) crawled along at best-case 8.3-Mbyte/sec rates.

Industry-blessed and serial-LVDS (low-voltage differential-signaling)-based Fibre Channel, whose development began in 1988 and whose specifications ANSI approved in 1994, started out at 133 Mbps and now runs at speeds as high as 2.125 Gbps—that is, 200 Mbytes/sec. It supports both twisted-pair copper and optical-interconnect options; the optical-interconnect option works at distances as long as 10 km. LVDS also supports both switched and arbitrated point-to-point and loop topologies. You can map a variety of communication protocols onto Fibre Channel fabric, including SCSI (the most common), IP (Internet Protocol), the AAL5 (ATM adaptation layer for computer data), link encapsulation, and IEEE 802.2 logical-link control.

Fibre Channel's performance and fea-

AT A GLANCE

- ▶ Parallel storage interfaces are running out of gas.

- ▶ SATA (Serial ATA) and SAS (Serial Attached SCSI) follow in the trail that Fibre Channel and SSA (Serial Storage Architecture) first blazed more than a decade ago.

- ▶ The SATA-bandwidth road map extends to at least 6 Gbps, and functional add-ons target multidisk configurations.

- ▶ SAS builds on a SATA foundation with enterprise-tuned enhancements.

ture robustness, however, translate to high silicon and software costs that have limited its use to high-end server SANs (storage-area networks). On the other end of the cost-versus-capability pendulum, PATA peaked at 100-Mbyte/sec speeds and, along the way, incorporated DMA features to reduce its load on the host processor. A 133-Mbyte/sec PATA version that Maxtor promoted received neither Intel's nor, consequently, most other drive manufacturers' blessings. PATA's developers designed it for mainstream systems with only two drives per channel. Its single-ended parallel-bus approach became increasingly constraining as data-transfer speeds accelerated and the need for adequate system airflow transformed from "nice to have" to "critical."

The master- and slave-storage peripherals on each PATA channel must contend for the available bandwidth, and the limitations of the half-duplex bus mean that PATA cannot perform simultaneous reads from one peripheral and writes to another. Installing two peripherals supporting different ATA peak transfer rates on the same channel invariably throttles

back the entire channel to the slower peripheral's speed. PATA's skew, crosstalk, and ground-bounce hazards will be familiar issues to anyone who's worked with a high-speed parallel bus employing a separate clock signal, and those issues also led to 18-in. maximum cable lengths. The large number of PATA signals translated to wide, bulky cabling; crosstalk concerns motivated many PC manufacturers to shy away from the narrower, rounded, and airflow-friendly PATA cables that have recently become popular with the system "modding" crowd. PATA's high signal count was also an issue for host- and peripheral-side controller-chip manufacturers, as was its reliance on advanced semiconductor-process-incompatible, 5V signaling.

The high-end Ultra320 SCSI, as its name implies, runs at 320-Mbyte/sec burst speeds. Its use of differential signaling, along with the embrace of advanced signal-optimization techniques, such as precompensation and active filtering, are key to the fact that SCSI performs better than PATA. Cable as long as 6m single-ended or 25m differential greatly extended SCSI's reach and made external cabling feasible. Until recently, when USB and FireWire took over the lead, SCSI was the predominant means of tethering not only external storage drives and arrays, but also high-speed peripherals, such as scanners, to PCs. Toward this end, SCSI-peripheral "chains" could have as many as eight or 16 devices.

Tagged-command queuing enables the drive to reorder its processing of incoming operation requests for compatibility with the data pattern on the rotating disk platter (**Figure 1**). This feature maximizes SCSI bandwidth efficiency. This reordering is particularly valuable when you consider the highly random, mixed

SCSI-DERIVED TRANSFORMATIONS

Although SAS's (Serial Attached SCSI's) impending fortunes aren't crystal-clear in the face of SATA's (Serial Advanced Technology Attachment's) onslaught, the SCSI command set has a far rosier future. SCSI is the most common protocol that Fibre Channel carries, whereas iSCSI (Internet SCSI) converts from that block-based SAN (storage-area-network) interconnect to

much cheaper but also potentially lower performance Gigabit or 10-Gbit Ethernet (**Reference A**).

Even cheaper file-based NAS (network-attached-storage) hardware based on protocols such as CIFS and NFS, however, may slowly but surely eat away at both of the SCSI-based SAN fabrics. Ximeta's NetDisks, NAS derivatives that the company calls NDAS (network-direct-

attached-storage) devices, require that at least one computer on the LAN have the necessary client software installed to communicate with the drive. Currently available software includes Windows, Mac OS, and Linux variants.

Other LAN clients can either directly interact with the NetDisk through their own installed software stacks or via another com-

puter through file sharing. D-Link's approach, on its Central Home Drives, requires no special client software but relies on Web-browser-based configuration utilities to create, delete, and edit directories and accomplish other disk-management tasks.

REFERENCE

A. Cravotta, Nicholas, "Fast track," *EDN*, Feb 20, 2003, pg 42.

read/write-access profiles that server-storage subsystems commonly see. SCSI shares PATA's high-signal-count shortcomings, though, and its added complexity makes it more expensive to implement than PATA. Any attempt to evolutionarily extend the interface beyond 320-Mbyte/sec speeds would further bloat this cost; the time is ripe for a revolutionary alternative approach.

Intel conceptually unveiled SATA at its February 2000 Spring IDF (Intel Developer Forum). (Version 1.0 of the specification arrived 1.5 years later at the Fall IDF.) Although SATA is command-set-backward-compatible with PATA, enabling you to continue running legacy software on it, its electrical and mechanical characteristics tackle many of the previously mentioned parallel-bus limitations. SATA migrates from a multiperipheral chain to a series of point-to-point links between the controller and each peripheral. At first, this transformation might seem to be a step backward from a flexibility standpoint. Consider, though, that each peripheral now has exclusive access to the 1.5-Gbps (150-Mbyte/sec), spread-spectrum-cognizant communication link and that each thin SATA cable contains only four active pins. Each link encompasses transmitting and receiving differential pairs, plus three grounds and a separate power connector.

The PHY of Fibre Channel and that of

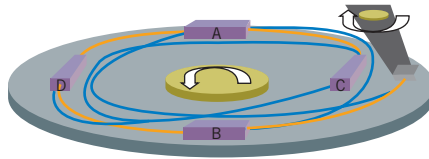


Figure 1 Command queuing enables this drive to reorder an incoming A-B-C-D command sequence as B-D-A-C to optimize the sequence for its organization and angular position, reducing what would otherwise be a 2.5-revolution task to a faster one-revolution alternative (courtesy Intel).

SATA share a high degree of commonality. They both employ LVDS, thereby removing PATA's 5V dependence. They also both support 8B/10B encoding that embeds the clock within the transmitted and received data and prevents accumulated dc offset, which would result from a disproportionate number of ones and zeros. The maximum SATA-cable length is 1m, enabling both internal and short, external peripheral connections (see sidebar "External-storage revolutions"). Second-generation, 3-Gbps SATA chips are now available from several suppliers, predating the PHY specification's imminent approval.

SATA will likely achieve rapid, widespread adoption in both notebook PCs and consumer-electronics applications. Ironically, the interface's high transfer rate is in neither case an important feature. Mobile computers frequently use slower hard drives than do desktop PCs to maximize battery life, and consumer-electronics gear also employs low-speed

drives to minimize cost, heat, and noise. However, Intel and its core-logic competitors want to as quickly as possible reduce their chips' storage-interface-signal count and discard that interface's 5V-tolerance requirement. Consumer-electronics manufacturers will similarly be attracted to the volume-cost efficiencies that the burgeoning SATA-drive supply delivers. Until optical drives, whose sustained read and write speeds are even lower than those of

their magnetic-drive counterparts, integrate native SATA interfaces, the systems will need to include SATA-to-PATA or PCI-to-PATA bridge chips somewhere along the link between controller and drive.

In desktop PCs, the PATA-to-SATA transition will take longer, but it will inevitably wrap up within the next few years (Figure 2a). Unlike with notebook PCs, in this case, users will for some time still want to add PATA-equipped hard drives and optical drives they already own to their new SATA-enhanced systems. PATA controllers integrated in core logic will fairly quickly disappear, though, and the same kinds of external bridge chips that notebook PCs incorporate will replace them. SATA controllers themselves are undergoing enhancement and standardization as part of Intel's AHCI (Advanced Host Controller Interface) initiative, which, among other things, improves SATA's power

EXTERNAL-STORAGE REVOLUTIONS

USB 2.0 and, to a lesser degree, IEEE 1394 ports are abundant in modern PCs, as are external magnetic and optical drives supporting these interfaces. Why, then, are Addonics Technologies and several other companies advocating external SATA (Serial Advanced Technology Attachment) adapters, drives, enclosures, and cabling? In a word: *speed*. Whereas USB 2.0 runs at a 480-Mbps link rate and first-generation IEEE 1394 clocks in at 400 Mbps, first-generation SATA delivers 1.5 Gbps. And, because SATA is a storage-tuned interface, its proponents claim that you'll obtain more efficient

use of its peak transfer capability in typical usage environments than you will with the higher overhead, more general-purpose USB and FireWire alternatives.

Addonics' benchmarks of PATA (parallel ATA) versus SATA on a common 120-Gbyte, 7200-rpm Western Digital ATA-100 drive with a 2-Mbyte cache, allude to the company's claims, although they don't directly compare SATA with USB 2.0 and FireWire (Figure A). Note that the performance disparity is greatest with buffered and sequential

operations that most efficiently access the hard drive's cache and platters. Look for additional companies, hungry for profitable new

markets, in the future to tout external SATA connectivity; SAS's longer cabling also makes it a natural fit here.

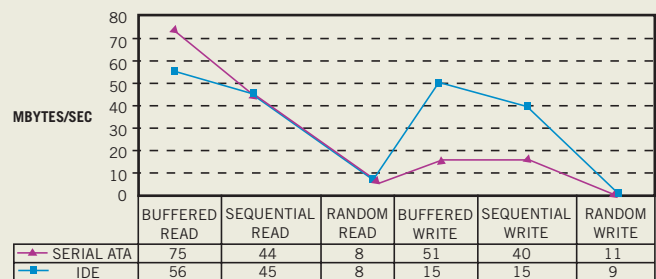


Figure A By tethering a PATA drive to an SATA interface via a bridge adapter, you might boost its performance in some access modes (courtesy Addonics Technologies).

management by eliminating the need for periodic active system polling for storage-peripheral insertion or removal, along with abolishing legacy master/slave peripheral and primary/secondary-channel limitations on the number of possible SATA peripherals in a system.

ABUNDANT OPTIONS, UNCLEAR OUTCOME

With the SATA specification's latest generation developments, the interface's proponents are revealing their intentions to migrate it upward into SCSI's traditional turf (Figure 2b). Along with the previously mentioned 3-Gbps (300-Mbyte/sec) PHY variant, SATA is incorporating support for native command queuing, conceptually similar to SCSI's tagged-command queuing. (Also, don't forget about the 6-Gbps upgrade lurking on the SATA road map.) The SATA Working Group has also provided an answer for those of you who are questioning any need for a high-speed storage interface, because only the transfers to and from any single drive's small RAM buffer can keep pace with it (Figure 3). Port-multiplier capability enables you to connect as many as 15 peripherals to each SATA link, amalgamating read and write traffic.

Addressing enterprise applications' concerns about data reliability and 24/7, 99.99% system uptime, SATA's port-selector feature provides for failover redundancy that switches from one storage peripheral to another in case of failure. And, to give warning of impending failures and more generally monitor storage subsystem operating conditions, SATA's

enclosure services controllers, such as those that Broadcom's ServerWorks subsidiary introduced a year ago, connect to the SATA link just as a storage peripheral would and then communicate with each peripheral in the storage array over I²C connections. The SATA Working Group is also defining PHY variants with wider output-voltage swings that enable driving longer backplane buses.

SCSI advocates claim that, although ATA's embrace of a high-speed serial bus is a solid

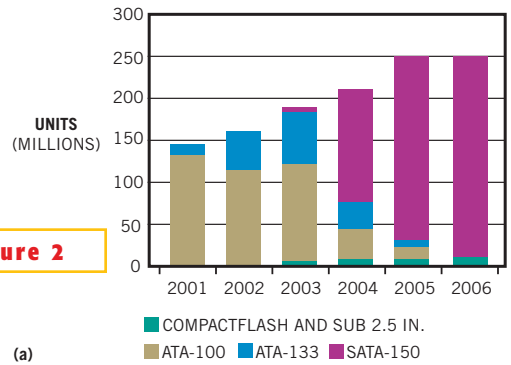
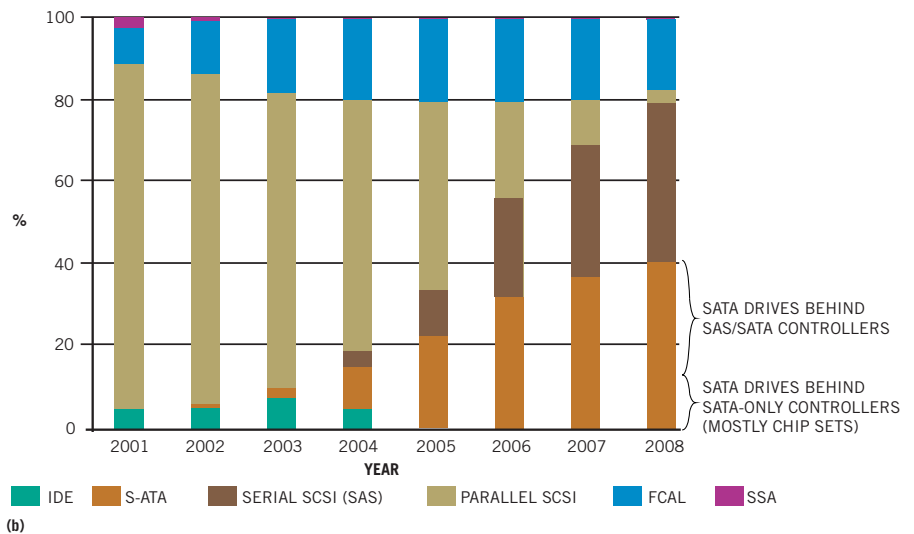


Figure 2



Serial-storage interfaces should rapidly displace parallel predecessors in both single-drive (a) and multidrive (b)—that is, enterprise, applications (courtesy Silicon Image and LSI Logic, respectively).

FOR MORE INFORMATION...

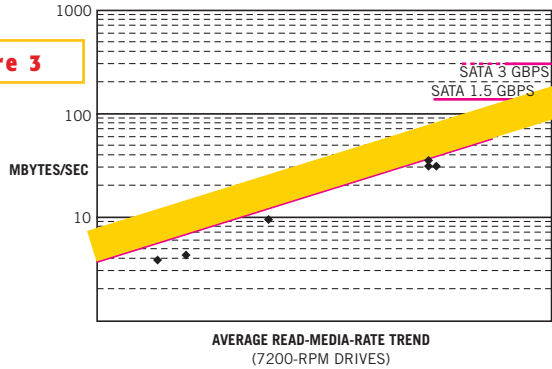
- | | | | | | |
|--|---|--|--|--|--|
| Acer Laboratories
www.ali.com.tw | ATI Technologies
www.ati.com | Hitachi
www.hitachi.com | Nvidia
www.nvidia.com | Rambus
www.rambus.com | 3ware
www.3ware.com |
| Adaptec
www.adaptec.com | Broadcom
www.broadcom.com | Intel
www.intel.com | Palmchip
www.palmchip.com | Samsung
www.samsung.com | Toshiba
www.toshiba.com |
| Addonics Technologies
www.addonics.com | Cornice
www.corniceco.com | LSI Logic
www.lsillogic.com | Plextor
www.plextor.com | Seagate
www.seagate.com | Via Technologies
www.viatech.com |
| Advanced Micro Devices (AMD)
www.amd.com | D-Link
www.dlink.com | Marvell Technology Group
www.marvell.com | PMC-Sierra
www.pmc-sierra.com | Silicon Image
www.siliconimage.com | Vitesse Semiconductor
www.vitesse.com |
| Agere Systems
www.agree.com | Fujitsu
www.fujitsu.com | Maxtor
www.maxtor.com | Promise Technology
www.promise.com | Silicon Integrated Systems (SiS)
www.sis.com | Western Digital
www.westerndigital.com |
| Agilent Technologies
www.agilent.com | HighPoint Technologies
www.highpoint-tech.com | NetCell
www.netcell.com | RAIDCore
www.raidcore.net | Spike Technologies
www.spiketech.com | Ximeta
www.ximeta.com |

For more information on the ever-evolving storage market, I encourage you to attend Intel's Developer Forums along with the pre-CES Storage Visions conference, and to regularly monitor Web sites such as the following: Storage Review at www.storagereview.com; Advanced Computer & Network's RAID.edu at www.acnc.com/04_00.html; the SATA Working Group at www.serialata.org; the SCSI Trade Association at www.scita.org and www.serialattachedscsi.org; and the T10 Technical Committee at www.t10.org.

start, even the latest generation SATA feature enhancements don't fully solve enterprise application problems. Pragmatically, SATA's lack of support for the SCSI command set is also a significant roadblock to adoption wherever continued use of legacy software is desirable. The SCSI advocates' response, SAS, builds on the SATA connector, cable, and PHY specifications, benefiting from SATA's volume-cost efficiencies, along with drafting off the SATA Working Group's development efforts (with several notable appendices). The SATA embrace extends to an acknowledgment that SATA drives' lower cost-per-gigabyte and higher densities might give them the overall edge versus SAS alternatives in some less critical enterprise applications, such as mirroring and backup.

To ensure storage flexibility, the SAS backplane and cable connectors are su-

Figure 3



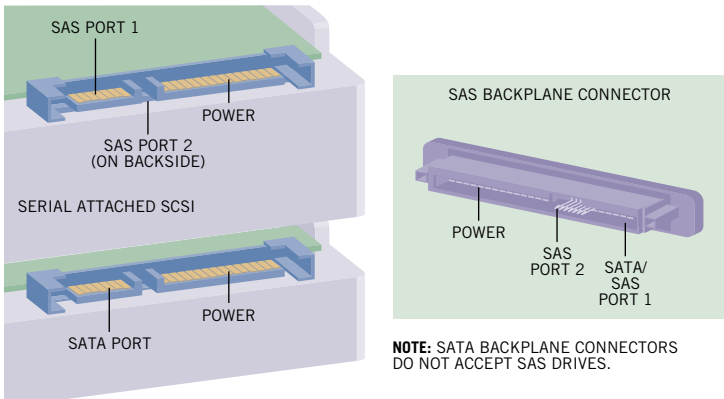
If you're wondering why the interface needs to run so much faster than a single drive can input or output data, SATA's bandwidth-concatenating port multiplier may provide an answer (courtesy Intel).

persets of the SATA versions with a second port for failover redundancy—not bandwidth multiplication (Figure 4). SAS links will transport both SCSI and ATA commands. ComSAS joins SATA's OOB (out-of-band) signals, ComReset (also known as ComInit) and ComWake, to enable communication and negotiation between initiator controllers and

target peripherals to determine whether each of those peripherals is SATA or SAS. Don't confuse the use of OOB here with its application in telephony, in which it refers to distinct wires intended for control signals—that is, physical OOB. With respect to storage, the same differential wires transmit and receive OOB signals and normal drive command and data traffic. OOB “chirps” have easily distinguishable lower frequency characteristics; out-of-tolerance PHYs can encode and decode these chirps, which are present only

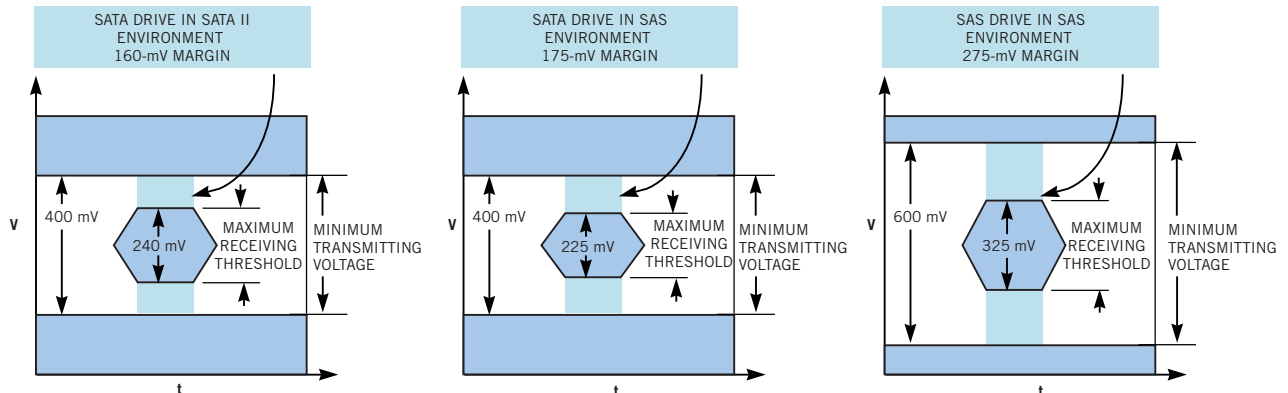
when the SATA or SAS link is quiescent. In this respect, the SATA and SAS implementations are of the “physical-in-band, logical-OOB” variety.

Production versions of first-generation SAS systems, due out this year, will run at the equivalent of second-generation SATA's 300-Mbyte/sec—thereby, similar to Ultra320—peak data rates, although their PHYs will be backward-compatible with 1.5-Gbps SATA peripherals. (This lower speed is the same speed at which many of today's SAS proof-of-concept prototypes, based on expensive FPGAs, run.) Unlike half-duplex PATA, SATA, and even parallel SCSI, SAS will be a full-duplex bus—that is, read and write traffic can concurrently run over the link—making each link's effective throughput twice that of SATA. With more complex interconnection fabric in mind, SAS is not only multitarget, like SATA when a port multiplier is present, but also multi-initiator in nature. SAS PHYs have wider output-voltage swings



SAS-backplane and -cable connectors accept both SAS and SATA drives (courtesy Intel).

Figure 4



NOTES: 1. ALL SIGNAL LEVELS ARE AT 1.5-GBPS SPEEDS.
2. AT 3.5 GBPS, SAS MARGIN IS 525 mV.

Figure 5

Higher-margin SAS levels enable longer cable and backplane spans (courtesy Intel).

than those of SATA, befitting their use in longer backplane and “near-enclosure” cabling configurations (Figure 5).

As for Fibre Channel, 4.25-Gbps-based standardization is complete, and chips are beginning to emerge from stalwarts such as PMC-Sierra. Fibre Channel’s cable length and near-term speed superiority will for the foreseeable future continue to secure it a defensible market niche versus SATA and SAS. Other SAN-interconnection alternatives, though, along with lower cost NAS gear, loom as more serious long-term competitive threats to it (see sidebar “SCSI-derived transformations”). The Fibre Channel technology road map currently shows a 10-Gbps variant scheduled to emerge during this decade, but 10-Gbit Ethernet may pre-empt and supplant it.

Keep in mind as you evaluate various serial-storage-interface options that these buses are only the highways upon which the command and data traffic flows. As an analogy, the speed capability of Germany’s Autobahn will largely go unrealized when you drive my 1.9-liter air-cooled-engine-powered ‘81 Volkswagen Vanagon camper, which sports a “0 to 55 in 11 minutes” bumper sticker. Similarly, robust storage interfaces depend on equally robust storage peripherals to achieve their full performance potential.

Historically, SCSI drives have offered higher random-access performance than their ATA brethren. SCSI drives’ high prices enable manufacturers to equip them with faster platters, larger RAM buffers, and more exotic controllers, along with access algorithms optimally tuned to servers’ access patterns. Developers have also designed and tested SCSI drives for the harsh environments and long operating lives that define the enterprise.

However, SCSI drives’ densities significantly lag behind ATA counterparts. Because SCSI drives spin faster, they draw higher power and, as a result, generate more heat. One fundamental way to keep power and temperature at reasonable levels is to constrain the SCSI drive’s per-platter radius. (Fujitsu, for example, claims that power consumption is proportional to the *cube* of the drive’s platter radius.) High reliability requirements also couple with high speed to increase the per-platter thickness to minimize the potential for cracking and shattering. Both the platter’s radius and its thickness factor in SCSI-drive-density shortcom-

ings; the radius limits the per-platter density, and the thickness limits the number of platters that a manufacturer can squeeze into the drive’s form factor.

The fact that workstations and servers often employ the drives in multistorage-peripheral arrays offsets SCSI-drive-density restrictions. When David Patterson and his fellow researchers at the University of California—Berkeley in the late 1980s initially envisioned RAID (redundant array of independent, or inexpensive, disks), they saw it as a way of boosting overall drive-subsystem performance, and, in speed-tailored RAID configurations, each drive incrementally adds to the overall subsystem density. Mirror- and parity-tailored RAID configurations, conversely, do *not* boost the storage subsystem’s size, though they do enhance its dependability.

Drives such as Western Digital’s Raptor 10,000-rpm SATA units, though, break through the somewhat-artificial historical price-versus-performance barrier between ATA and SCSI—a barrier that other, stronger historical promoters of SCSI, such as Maxtor and Seagate, probably wish would remain intact. In the process, drives such as the Raptor series could potentially transform the building blocks of storage subsystems. SATA’s low pin count couples with the affordability of modern disk drives to drive the proliferation of RAID into mainstream-PC systems. And, hungry for any and all MIPS-gobbling applications to feed their ever-faster CPUs, AMD and Intel are doing their utmost to ensure that host software rather than discrete hardware will control these RAID configurations. □

REFERENCE

1. Dipert, Brian, “Short-term detour, or long term bypass?” *EDN*, May 29, 2003, pg 30.

ACKNOWLEDGMENTS

Special thanks to Knut Grimsrud and Dave Dickstein from Intel for the excellent IDF materials that inspired this article.

AUTHOR’S BIOGRAPHY

Technical editor Brian Dipert is off to install PATA and SATA drives in several PCs he’s building, which he’ll be benchmarking and putting through their paces for the hands-on cover story in EDN’s March 4, 2004, issue. Reach him at 1-916-454-5242, fax 1-617-558-4470, bdipert@edn.com.