The Premier Advanced Recording Technology Forum

THIC Inc.

Disk for Real-time Data

Phil Brunelle
Conduant Corporation
1501 S. Sunset St., Suite C, Longmont, CO 80501
Phone:303-485-2721 FAX: 303-485-5104
E-mail: phil@conduant.com

Presented at the THIC Meeting at the Sony Auditorium, 3300 Zanker Rd, San Jose CA 95134-1940

March 4-5, 2003

Disk for Real-time Data

There's a Universe of Data out there……Just Waiting…

Topics

- *Whats and whys of Real-time Data?*
- *Why disk?*
- *Understanding the Architecture*
- *Specsmanship*
- *Performance graphs*
- *Summary*

Whats and Whys of Real-time Data?

What is Real-time Data?

- *Is a data stream produced by or representing a physical event with specific temporal characteristics*
- *While recording, a stream of data that must be recorded at a continuous rate to accurately capture the temporal nature of the data*
- *While playing back, a stream of data that must be reproduced at a specific continuous rate to accurately represent the temporal nature of the data*

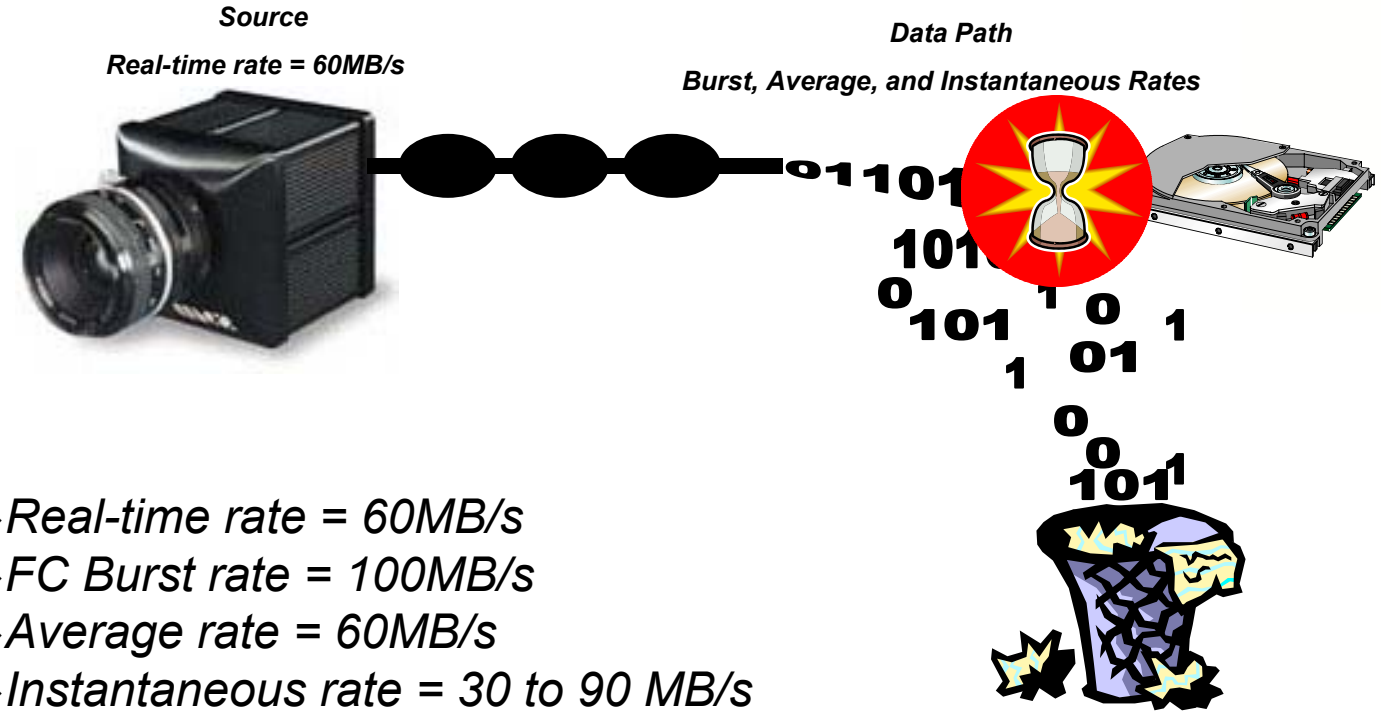
Why record and play back real-time data?

- *Why record?*
  - *To capture details of physical events for analysis with data processing techniques or through human examination*
  - *So that we can play it back in real-time*
- *Why play back?*
  - *To reproduce a physical event in the same manner that the event originally occurred*
  - *To produce a physical event with a computer-generated representation of real-time data (i.e. simulation)*

What is the difference between data processing and real-time storage?

- *Data Processing Storage:*
  - *Maximizes overall throughput of many small, unrelated transactions, rather than being most efficient on any one*
  - *Generally not concerned with variable access times for any individual transaction*
  - *Logically sequential data physically scatters through the course of normal operations*
  - *Requires exactness of data but NOT temporal repeatability*
- *Real-time Storage:*
  - *Sustains continuous throughput of a single transaction*
  - *Uses various techniques to compensate for the unpredictable mechanical delays inherent in magnetic storage devices*
  - *Data records and remains as a continuous linear record*
  - *Generally requires both exactness of data and temporal repeatability*

What is the Real-time storage problem?



- Real-time rate = 60MB/s
- FC Burst rate = 100MB/s
- Average rate = 60MB/s
- Instantaneous rate = 30 to 90 MB/s
- Data is lost during periods when instantaneous rate falls below 60MB/s even though average and burst rates are higher than 60MB/s

What are some real-time applications?

- Radar
- Imaging
- Stress Measurements
- Radio Astronomy
- Torpedo runs
- Entertainment
- Surveillance
- Virtual Reality
- Law Enforcement
- Digital Radio
- Missile tracking
- GPS
- Acoustics
- Code/logic/network trace
- Instrumentation
- Semiconductor Test
- Telemetry

Why Disk?

Why Disk?

- *Value*
- *Performance*
- *RAS (Reliability/Availability/Serviceability)*
- *Many packaging options*
- *Integration*
- *Environmental*
- *Archival*
- *Futures*

Why Disk?

- *Value*
  - *Low cost per gigabyte ($1.40/GB for IDE on 1/30/03)*
  - *Low cost per megabyte/second*
  - *Negligible cost to maintain*
  - *Low cost to upgrade*
  - *Minimal cost for on-site sparing*
  - *Reduced power, cooling, footprint/TB*
  - *Economical data mirror for redundancy*

Why Disk?

- *Performance:*
  - *Scalable, multiple drives can be ganged together sustain very high data rates (gigabits or 10s of gigabits/sec and beyond)*
  - *Fast random access during playback (10s of milliseconds to totally different locations)*
  - *Rapid offline data duplication, local or remote*
  - *High-speed real-time remote data mirror*
  - *Simultaneous continuous recording and random data review*

Why Disk?

- *Reliability/Availability/Serviceability*
  - *Low MTBF (< 1 in 10^{14} bits read) per drive*
  - *Ramp loaded heads*
  - *Good shock/vibration using 3-1/2" drives, better with 2-1/2" drives, and best with Flash disks or specifically designed packaging*
  - *Economical on-site sparing minimizes MTTR (mean time to repair)*
  - *Replacement drives available at store down the street or overnight from mail order*

Specifications based on HGST (formerly IBM) DeskStar 120.

Why Disk?

Many packaging options:

Why Disk?

- *Integration*
  - *Easily integrated with most common computer systems*
  - *Remote control and data access over Ethernet*
  - *Seamless “infinite” recording (TK200)*

Why Disk?

- *Environmental/Facilities*
  - *Low power and reduced dissipation (~10W/drive operational, less with 2-1/2" drives)*
  - *Reduced footprint (4.05GB/in^3 with 250GB drives, 2TB per TK200 removable drawer)*
  - *Reduced noise (~ 3 bels/drive)*

Specifications based on Hitachi (formerly IBM) DeskStar 120.

Why Disk?

- *Archival*
  - *Low cost per uncompressed gigabyte ($1.40/GB for IDE on 1/30/03)*
  - *High Density (4.05GB/in^3 with 250GB drives, 2TB per removable unit)*

Conduant TK200

Why Disk?

- *Futures:*
  - *It just keeps getting better, better, and better!*

Understanding the System Architecture

Not all recording systems are alike

- *Total system performance is dependent on:*
  - *Data organization on disk*
  - *Individual disk performance*
  - *Dedicated Disk/Data engine*
  - *Data path*
  - *Operating System/File System*
  - *Data buffer*
  - *DSA (Dynamic Storage Allocation)*

Understanding the Architecture

- *Data Organization*
  - *Sequential is the natural organization for both disk drives and real-time data*
  - *Sequential yields high performance, predictability, and repeatability*
  - *Fragmented or scattered is BAD!*
  - *OS/File System organized is fragmented!*

Understanding the Architecture

- *Disk Performance:*
  - *Bit rate between media and head (Mbits/sec)*
  - *Head switch and incremental seek delays*
  - *Overhead fields (Inter-sector gaps, ECC, PLL sync, etc.)*
  - *Head position from OD (Outer Diameter)*
  - *Maximum interface transfer rate*

Understanding the Architecture

- *Dedicated Disk/Data engine is GOOD!*
  - *Can ensure physical sequentiality*
  - *Predictable, repeatable, performance*
  - *Responds to asynchronous events at hardware speeds*
  - *Pushes drives to their inherent capabilities*
  - *Higher level of resource utilization*

Understanding the Architecture

- *Customer data rate defines the minimum instantaneous data rate*
- *Best data path*
  - *Point-to-point, single pass*
  - *Dedicated (unshared)*
  - *Simple protocol*

Understanding the Architecture

- *Operating System/File System not designed for real-time data*
  - *Responds to asynchronous events at software speeds*
  - *Unpredictable shared resource*
  - *Layers of abstraction produce logically sequential but physically scattered data*
  - *Drives operate at a fraction of their inherent capabilities*
  - *A compromise to meet the needs of all kinds of users and services*
  - *Storage destination committed when file is opened for writing*

Understanding the Architecture

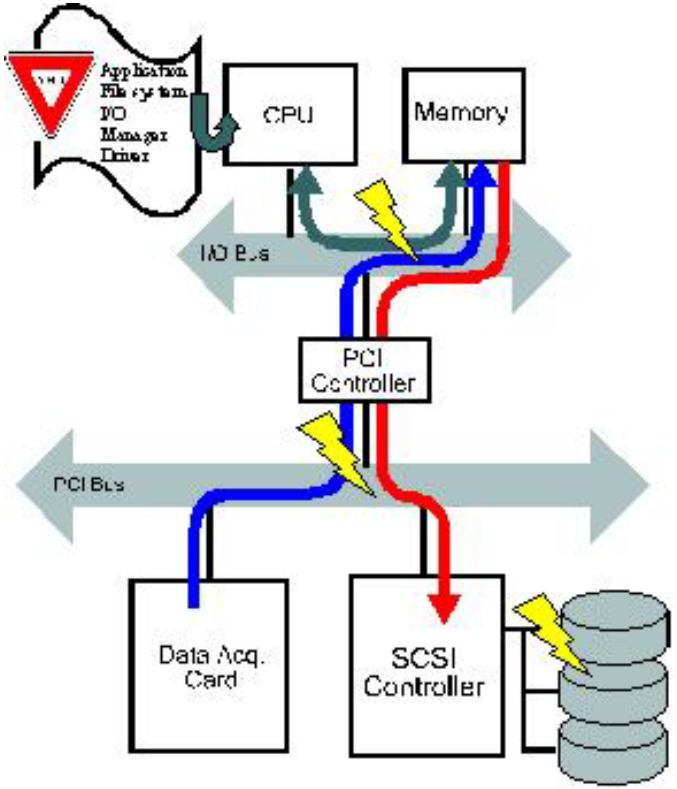
- *Data buffer*
  - *More is better*
  - *Think "number of seconds at a given data rate", not number of megabytes*
  - *Larger buffer allows more time for disk drive error recovery*

Understanding the Architecture

- *DSA (Dynamic Storage Allocation)*
  - *Priority on getting data to magnetic media above all*
  - *Compensates for underperforming drive with underutilized drive*
  - *Picks up where data buffer leaves off*

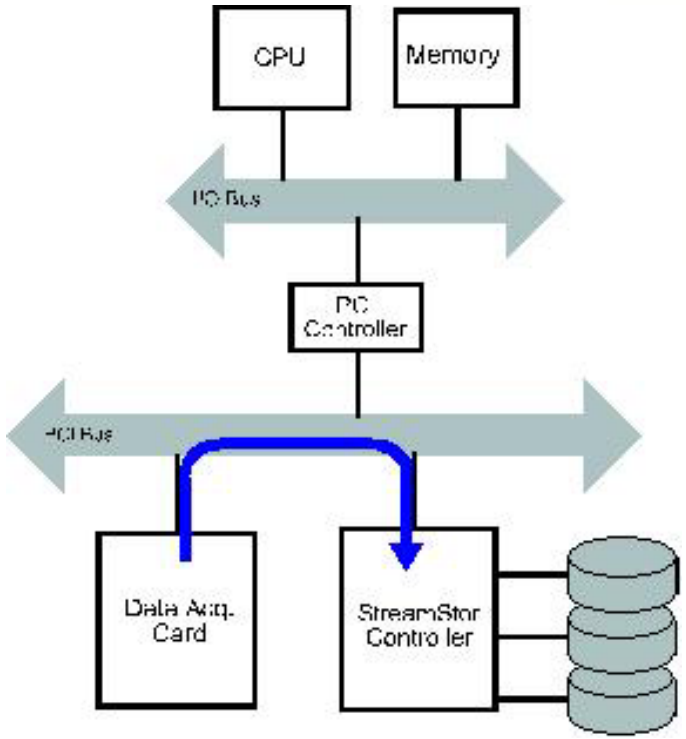
Simple solution: The first solution people think of is riddled with problems:

- *Same data moves twice on same PCI bus*
- *Same data is addressed twice in system memory*
- *Real-time data competes with processor instruction fetches*
- *Data moves from source to disk under control of application software*
- *Storage is under control of OS*
- *Data organized per OS file system is not physically sequentially*
- *Source card must include data buffer sufficient to account for when PCI bus is busy with other traffic*



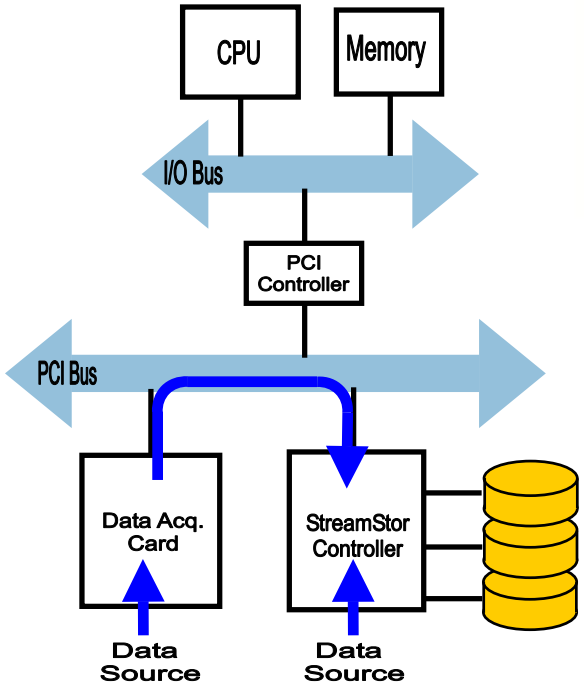
Card-to-card and dedicated real-time engine. A significant improvement:

- *Data moves only once on PCI bus, thus doubling possible bandwidth*
- *DMA engine on source card pushes data directly to storage controller with no software intervention*
- *Dedicated hardware engine for disk storage independently controls with primitive commands to ensure physically sequential recording*
- *Source card must include data buffer sufficient to compensate for when PCI bus is busy with other traffic*



Add a direct customer data path into the storage controller for the best solution.

- *Similar to card-to-card solution except that a dedicated data interface:*
  - *Reduces or eliminates the need for buffer FIFOs on the source card*
  - *Totally eliminates the possibility of computer interference in data path*
  - *Simplifies the interface protocol due to lack of required arbitration*



Specsmanship

Specsmanship

- *Watch out for "Magic" Numbers:*
  - *usually theoretically maximum burst rates that are not sustainable*
  - *40,80,160, and 320 MB/s for SCSI*
  - *100 and 200 MB/s for FC*
  - *Highly impressive for sales purposes, worthless for the real-time recorder application*

Specsmanship

- *Benchmark software:*
  - *Fine for relative comparisons of different storage systems under identical conditions Optimized to test raw disk performance*
  - *does NOT reflect performance across the complete data path*
  - *Does not reflect performance of application software*
  - *Full data path performance is generally much less than disk benchmark under OS*

Specsmanship

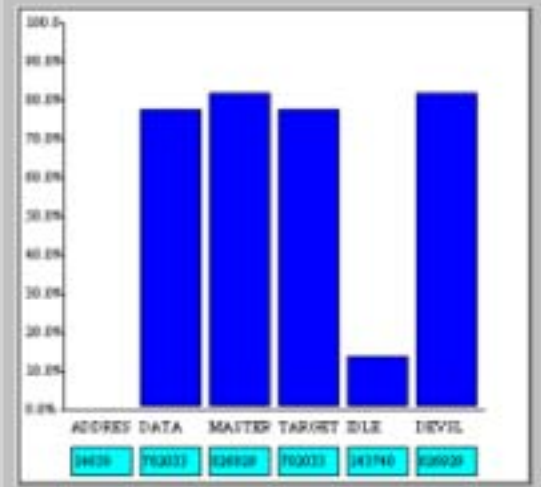
- ***Ask*** *the supplier:*
  - *How did he arrive at the minimum sustained data rates from data source to disk?*
  - *If he is being optimistic or conservative? Why?*
  - *If the performance is contingent on certain conditions?*
  - *To explain his architecture from data source to disk and back. Do you see any problems?*
  - *If his company or a strategic partner owns the IP? If not, he may not really know the details and he may not be able to directly address problems should they arise.*

Performance Graphs

Catalyst Enterprises, Inc. TA300 Exercises - Connected to LPT1

New File Edit View Run Record_Play Utilities Window Help

Bar Graph Window



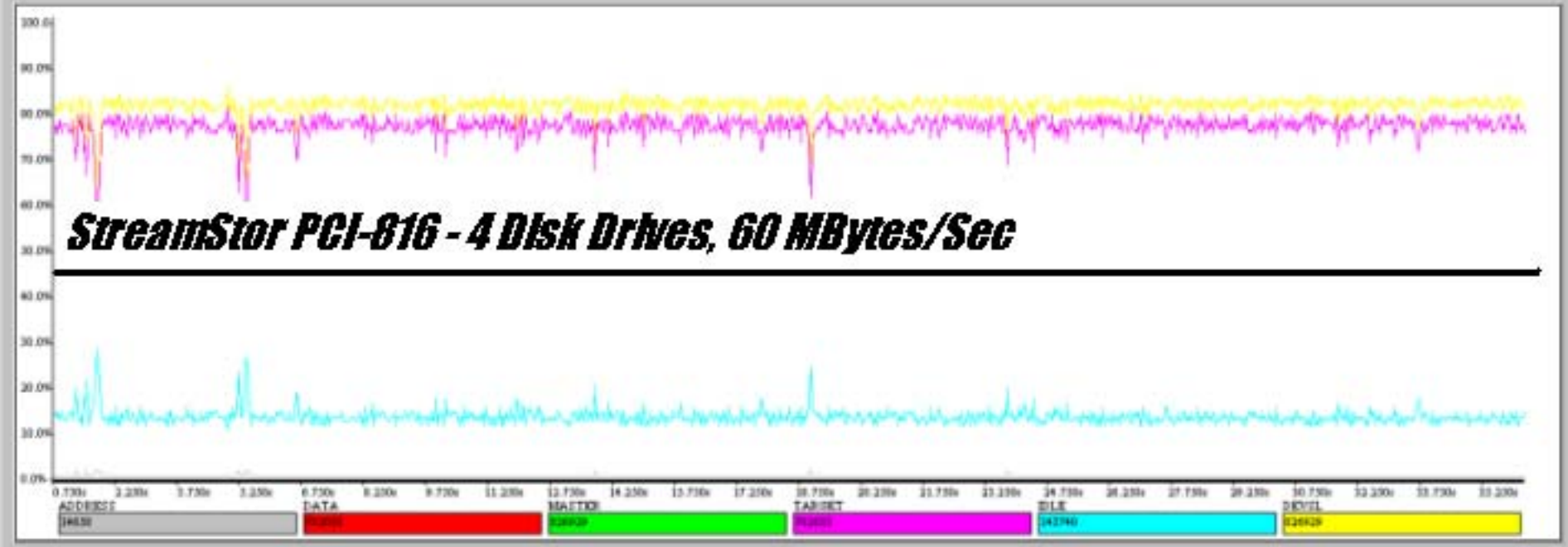
Reports Window

(32 bit) PCI Performance, Efficiencies and Throughput

Target_Eff = 94.571 %
Master_Eff = 100.000 %
XFER_Eff = 94.571 %
Avg_Burst = 53.352 Data_Per_Address
Data_Rate = 104.167 MBytes/Sec
Wait = 5.429 %
Bus_Occupied = 85.626 %_(Entire_Bus_only)
Idle = 14.374 %_(Entire_Bus_only)

ADDRES DATA MASTER TARGET IDLE DEVSL

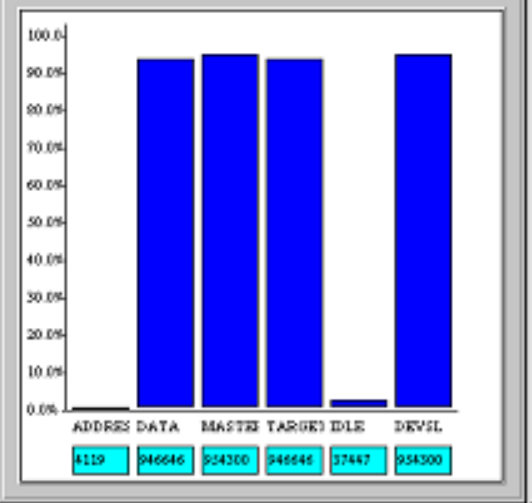
Histogram Window



Catalyst Enterprises, Inc. TA300 Exercizer - Connected to LPT1

New File Edit View Run Record_Play Utilities Window Help

Bar Graph Window



Reports Window

(32 bit) PCI Performance, Efficiencies and Throughput

Target_Eff = 99.198 %

Master_Eff = 100.000 %

XFER_Eff = 99.198 %

Avg_Burst = 229.824 Data_Per_Address

Data_Rate = 126.093 MByte/Sec

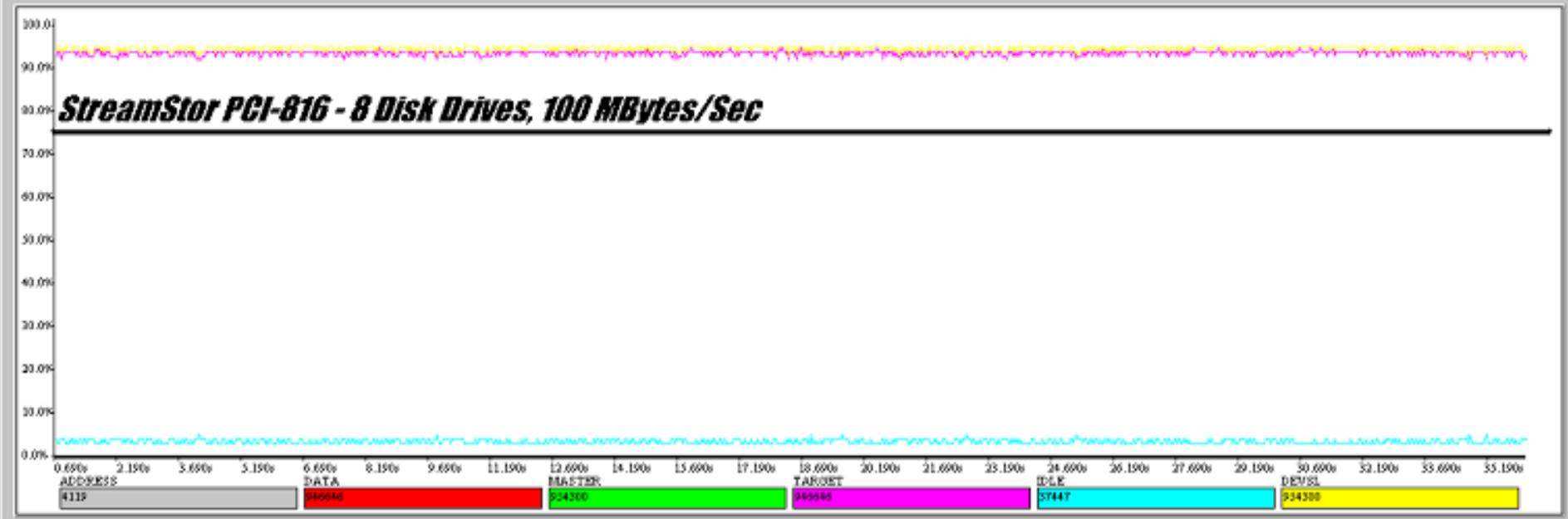
Wait = 0.802 %

Bus_Occupied = 96.255 %_(Entire_Bus_only)

Idle = 3.745 %_(Entire_Bus_only)

ADDRESS	DATA	MASTER	TARGET	IDLE	DEVSL
4119	946646	934300	946646	57447	934300

Histogram Window



CAPS NUM SCRL

Summary

Summary

- *Think **Minimum Instantaneous Data Rate**, NOT burst data rate and NOT average sustained data rate*
- *Real-time data can require the exactness of data processing data but definitely includes the temporal component*
- *Current IDE disk technology, when coupled with controller designed for real-time storage, fills most of the required and desired aspects for real-time recording and playback*
- *Understand the specifications, how they were arrived at, and what the seller will guarantee.*

There's a Universe of Data out there……Just Waiting…

Thank You!