LA-UR- 97-4905

Approved for public release; distribution is unlimited.

Title: HIPPI-6400 -- A Summary

CONF-9705247--

Author(s): Donald E. Tolmie

RECEIVED
APR 06 1998
OSTI

Submitted to: Ninth Annual Workshop on Interconnections Within High Speed Digital Systems, May 17-20, 1997, Santa Fe, NM

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

19980423 175

Los Alamos
NATIONAL LABORATORY

Form 836 (10/96)

HIPPI-6400 – A Summary

Don E. Tolmie
Los Alamos National Laboratory
MS-B25, Los Alamos, NM 87545, USA
Phone: (505) 667-5502
FAX: (505) 665-7793
E-mail: det@lanl.gov

ABSTRACT

HIPPI-6400, an emerging standard for LANs and system area networks (SANs), supports data rates of 6400 Mbit/s (800 MByte/s). HIPPI-6400 borrowed freely from ATM, Ethernet, and the original HIPPI. The HIPPI-6400 features and design choices are discussed.

HIPPI-6400 – Designing for speed

Don E. Tolmie
Los Alamos National Laboratory

The High-Performance Parallel Interface – 6400 Mbit/s (HIPPI-6400) draft standards are next generation LAN interconnect standards; beyond the current gigabit, and near gigabit, standards. Operating at 6400 Mbit/s (800 MByte/s), full-duplex, HIPPI-6400 ensures maximum compatibility with the Ethernet, Gigabit Ethernet, ATM, and HIPPI installed base. The original HIPPI standards, running at 800 and 1600 Mbit/s, developed and first deployed almost 10 years ago, pioneered higher speed interconnect technology. The original HIPPI, along with a proposal from Silicon Graphics Inc. provided the starting point for HIPPI-6400.

HIPPI-6400 is based on the best features of several successful interfaces – drawing from ATM, Ethernet and the original HIPPI specifications. From ATM it borrowed a small 32-byte micropacket (like a 48-byte ATM cell), and four Virtual Circuits (fewer than ATM, but limited for performance reasons). From Ethernet it borrowed the MAC header to allow easy translation to other popular protocols, and to use existing Ethernet-based control and management tools. From the original HIPPI it borrowed the large message size capability, credit-based flow control, encoding scheme for dc-balance, and a cable using multiple twisted-pairs (or optical fibers), for the data path. Features of HIPPI-6400 not found in any of these interfaces include end-to-end as well as link-level checksums, automatic retransmission at the physical layer to correct flawed data, and a data rate of 6400 Mbit/s. As in other gigabit technologies, HIPPI-6400 systems will be switched rather than have multiple devices sharing a common bus or medium.

The HIPPI-6400 standards are being developed in ANSI Task Group T11.1 (see the web page at **http://www.cic-5.lanl.gov/~det** for meeting notices, meeting minutes, and draft documents). In relation to the OSI Reference Model, HIPPI-6400-PH (Physical Layer) specifies the physical and data link layers. HIPPI-6400-SC (Physical Switch Control) specifies a network Layer for controlling physical layer switches. T11.1 completed their work on these documents in October 1997, and forwarded them for further review and balloting. The HIPPI-6400-PH and -SC documents are expected to complete their processing and become approved ANSI standards in late 1998. In addition, Task Group T11.1 is working on a transport layer standard, initially part of HIPPI-6400-PH, called the Scheduled Transfer Protocol (ST). Scheduled Transfer takes advantage of the high-speed reliable HIPPI-6400 lower layers, and provides additional performance by bypassing parts of a host's operating system. Scheduled Transfer specifies mappings for use on HIPPI-6400, Ethernet, ATM, and Fibre Channel.

System features –

– physical links are point-to-point, bi-directional, symmetrical

– logical links are simplex (transmit and receive are completely separate)

– user data rate of up to 6400 Mbit/s (800 MByte/s) in both directions simultaneously

– the 6400 Mbit/s data rate does not include control information (i.e., 99.6% is available to the user)

– control information is transferred in parallel with the data (i.e., out of band)

– links (physical-layer) are defined in HIPPI-6400-PH

– physical-layer switches are defined in HIPPI-6400-SC

– first installations in 1998

Micropackets –

– the basic transfer unit on a link

– a micropacket contains 32 data bytes and 8 control bytes (all transferred in 40 ns)

– the small transfer unit (i.e., a micropacket), results in low latency for short messages and a component for large transfers

– a hop-by-hop credit based flow control at the micropacket level prevents data overrun

– micropackets use a 16-bit CRC for link-level checking

– flawed or lost micropackets are automatically retransmitted, providing in-order, reliable, data delivery

– retransmission is based on time outs (for a robust system)

– null, or credit-only, micropackets are transmitted when there is no data to send

Virtual Channels –

- 4 Virtual Channels (VC0–VC3) in each direction on each link (supports multiplexing)
- only 4 VCs as opposed to ATM's many (for performance and buffering reasons)
- VCs are assigned to specific message sizes, e.g., VC0 ≤ 2176 bytes, VC3 ≤ 4 GBytes
- credits are separate for each VC (congestion on one VC doesn't affect the other VCs)
- small messages on VC0 will not be blocked behind a large message on another VC
- all of the micropackets of a message are delivered in-order on a single VC
- the VC number is not changed within a HIPPI-6400 fabric

Messages –

- messages are a sequence of micropackets
- messages carry the user's data payload
- messages use an Ethernet MAC header for routing (48-bit Source and Destination addresses, 32-bit length)
- header compatibility with Ethernet makes translation to other media easy (e.g., Gigabit Ethernet, ATM, etc.)
- a 32-bit message length parameter allows very large messages
- messages use the IEEE 802.2 LLC/SNAP header for upper-layer protocol selection based on EtherType
- messages use an additional 16-bit CRC for end-to-end checking
- 16-bit end-to-end CRC is checked at each node along the path
- a Tail bit marks the end of a message
- retransmission of flawed messages is the upper-layer protocol's responsibility
- no undetected errors unless ≥ 6 bits in a micropacket are in error

Signaling characteristics –

- parallel signals strobed by a common CLOCK signal
- parallel architecture allows each signal to run at 0.5 or 1.0 GBaud, vs. about 10 GBaud if serialized
- 4B/5B encoding for ac coupling and dc balance (simpler than 8B/10B to implement, and 20 copies are on an interface chip)
- up to 10 ns differential skew allowed between signals (compensates for wire length differences and temperature)
- dynamic skew adjustment uses 40 ns every 10 μs (eats 0.4% of the bandwidth)

Physical attributes –

- SGI's protocol chip, called SuMAC (Super-HIPPI MAC), is implemented in CMOS
- the protocol is tuned for distances up to 1 km (without speed degradation)
- distance limitation was based on on-chip receive buffer size
- distance limitation is now built into some of the protocol field sizes
- ac-coupled 23-pair copper cable interface for distances up to 40 m (each signal is 500 MBaud max., uses a 100-pin connector)
- 12-fiber fiber optic ribbon for longer distances (up to 10 km) is being defined in a separate standards document (each signal is 1 GBaud max., uses standard MTO connector)
- 850 nm optics with multimode fiber, up to 300 m (problems with eye-safety due to multiple fibers in the ribbon)
- 1300 nm optics with either multimode fiber (up to 300 m) or single mode fiber (up to 10 km)

Acknowledgements

The Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. The author's work was performed under auspices of the U.S. Department of Energy. Silicon Graphics Inc., with Dr. Greg Chesson leading their efforts, has contributed the majority of the HIPPI-6400 technical innovations. The HIPPI standards committee, with participation from many people throughout the industry, has worked tirelessly to document HIPPI-6400, and Scheduled Transfer, as ANSI standards.

Biography

Don Tolmie joined the Los Alamos National Laboratory in 1959, and has been involved with networking of supercomputers since 1972. He has led the HIPPI standards efforts since HIPPI's conception in 1987, and is presently the Chairman of T11.1. He holds a BSEE from New Mexico State University (1959), and an MSEE from University of California - Berkeley (1961).

M98004354

Report Number (14) LA-UR--97-4905

CONF-9705247--

Publ. Date (11) 199705

Sponsor Code (18) DOE/DP, XF

UC Category (19) UC-705, DOE/ER

DOE