PARAMNet-3 Packet Routing Switch : ANVAY

- Modular solution supporting 8→48 ports
- Each port supporting 10 Gbps, full duplex communication over CX-4 cables
- Interval labelling based packet routing
- Wormhole routing for reducing latency
- 32Kbyte (ingress) and 16 Kbytes (egress) packet buffers
- Near neighbour communication using dedicated communication path
- Port to port latency < 2 μsec
- Near wire speed performance (1.1 GByte/sec per port)
- Pause/Resume based flow control with back pressure
- Packet payloads up to 4KBytes
- Support for unicast and broadcast
- Fully manageable over LAN and RS232
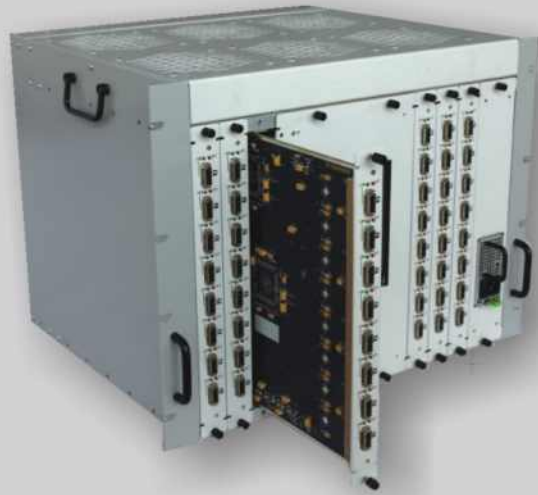
KSHIPRA Software Stack

- Linux support
- 64k end points
- RDMA centric architecture
- uDAPL light weight protocol with kernel bypass
- OpenIB adaptation with IPoIB, SDP and iSER protocols
- Requires no modification in the existing TCP based applications
- Communication libraries MVAPICH, Intel MPI supported
- Targeted for HPC, Storage/Database and enterprise applications

HTDG — HARDWARE TECHNOLOGY DEVELOPMENT GROUP
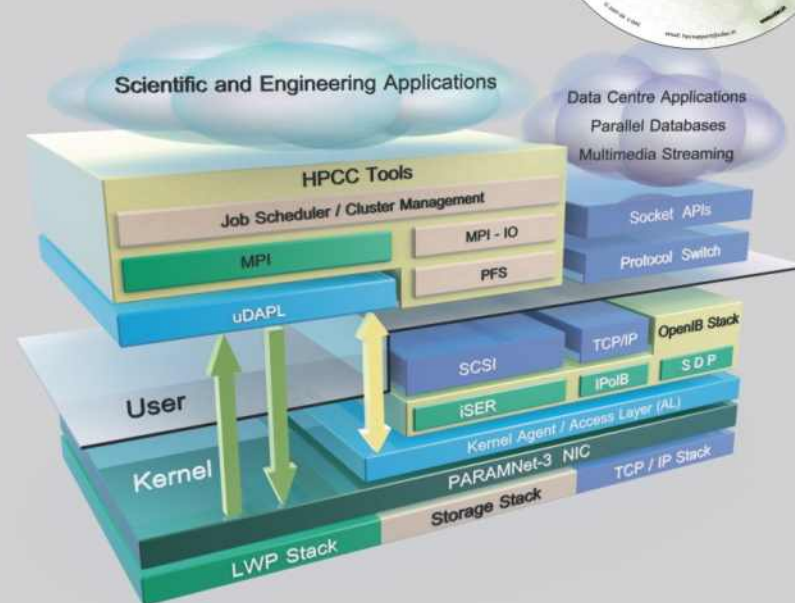
CDAC

PARAMNet-3

High Performance Cluster Interconnect

C-DAC/HTDG PARAMNETBro'Mar'09

CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

www.cdac.in

PARAMNet-3 is a high performance cluster interconnect developed indigenously by C-DAC. With this development, C-DAC has joined an elite group of system developers worldwide, capable of providing a critical high performance networking component for building supercomputing systems.

The main application of PARAMNet-3 is, as a primary interconnect for PARAM Yuva. Other application areas identified for its deployment are storage and database applications. It is also an integral component of other HPC solutions offered by C-DAC.

PARAMNet-3 consists of tightly integrated hardware and software components. The hardware components consist of Network Interface Cards (NIC) based on C-DAC's fourth generation communication co-processor "GEMINI", and modular 48-port Packet Routing Switch "ANVAY". The software component "KSHIPRA" is a lightweight protocol stack designed to exploit capabilities of hardware and to provide industry standard interfaces to the applications.
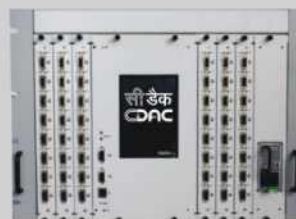
Design Approach

Performance and scalability of a supercomputing cluster is largely determined by the interconnect characteristics. PARAMNet-3 offers several advantages over commodity, off the shelf LAN solutions by taking the following approach:

‣ Offload the transport layer functionality to hardware, by developing specialized "communication co-processor" hardware (CCP). Such a processor effectively decouples communication tasks from host processor, providing much better performance and freeing the host processor for running user application. A packet processor based solution also allows for near wire speed useable bandwidth and low latency path for faster communication.

‣ Provide a extremely low latency, high throughput network cloud by developing specialized packet routing switches.

‣ Provide a standard set of application level interfaces, so that the user applications can be made to run on the specialized hardware with minimal changes.

‣ Provide legacy support for TCP/IP based applications.
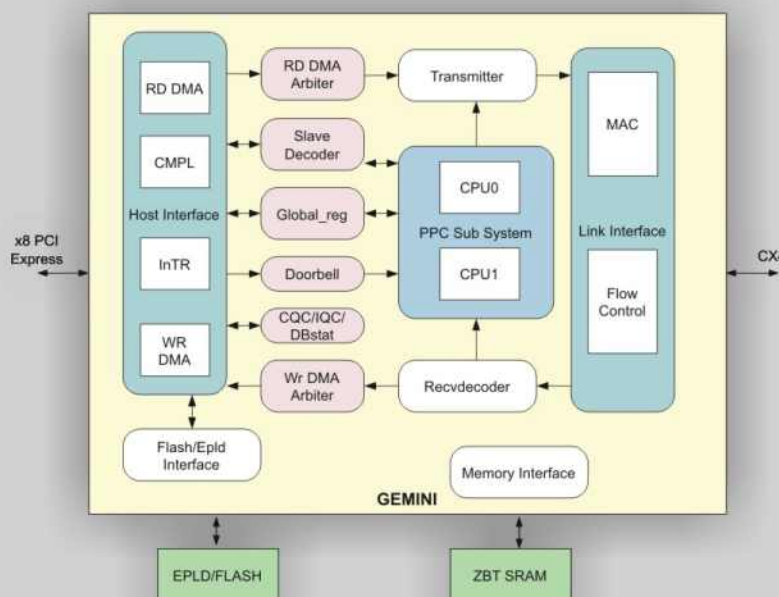
Communication Co-processor : Gemini

The Gemini Co-processor is fine tuned for HPC applications over uDAPL and OpenIB software stacks. It offloads transport layer functionality to hardware for supporting send / receive, Remote Direct Memory Access (RDMA) and Datagram oriented communication protocols, allowing compute nodes to communicate efficiently. Direct user level hardware access to 4k simultaneous connection oriented and connectionless hardware endpoints (scalable upto 16k) is provided to support applications of larger problem sizes. Gemini is capable of performing I/O from paged virtual memory with byte aligned addresses and reliable communication of maximum 4 GByte message length.

Gemini Architecture

Packet Routing Switch : Anvay

The modular, 48-port packet routing switch 'Anvay' provides a near wire speed (10Gbps) packet routing capability with very low latencies. Anvay switch is based on Spider Line Card and Backplane subsystems. Each spider card support up to eight ports, with each port working at 10Gbps, full duplex speeds. A fully populated switch supports approximately 1 Terabits / second of aggregate system throughput. The switch supports multi-level switching, allowing cascading of switches for supporting large clusters.
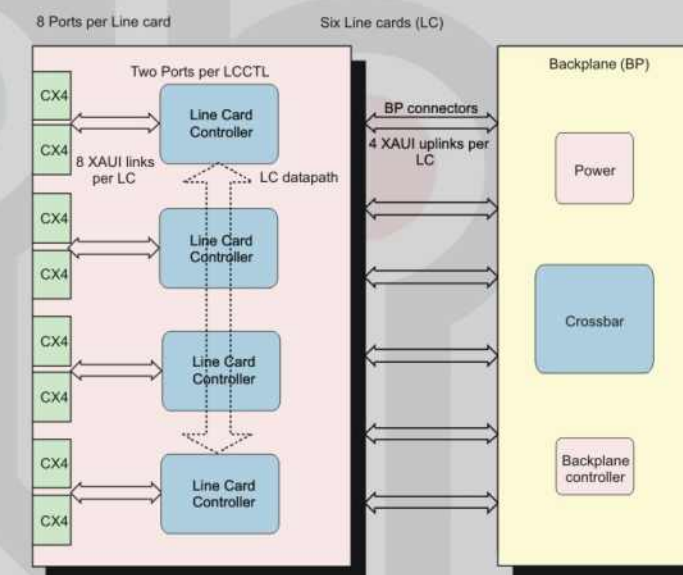
Communication Substrate : KSHIPRA

KSHIPRA is a scalable communication substrate over PARAMNet-3. Its a software environment that enables applications to communicate faster using RDMA. KSHIPRA allows the HPC, storage and enterprise applications to take advantage of PARAMNet 3's low latency, high bandwidth capability and drastically cuts down the CPU usage, by using DAPL transport mechanism. It also depicts PARAMNet-3 as a standard IP network. Thus all TCP/IP applications can be run without any modifications. MPI, SDP, IPoIB and iSER are included in the protocol set.

KSHIPRA, available as a bundled CD is a complete environment consisting of drivers, libraries, cluster management tools and documentation.

Specifications

PARAMNet-3 Network Interface Card (NIC)

‣ Based on GEMINI communication co-processor

‣ PCI express (x4/x8) full duplex based host interface

‣ 10Gbps Full duplex CX-4 link interface

‣ 16 MBytes onboard memory

‣ Transport protocol offload engine supporting send / receive, RDMA and Datagram protocols

‣ Proprietary hardware protocol implemented using Dual 32-bit CPUs for low latencies and reliable communication

‣ Direct user level access to 4k simultaneous connection oriented or connectionless hardware endpoints (scalable upto 16k)

‣ Capable of performing I/O from paged virtual memory with byte aligned addresses

‣ Address Translation table for VA-PA with protection implemented in hardware

‣ Communicates to host through Interrupt and Completion Queues