Impact of huge bandwidth optical interconnection network and a new memory paradigm with global address space for BDEC systems

Tomohiro Kudoh, Shu Namiki, Ryousei Takano, Kiyo Ishii, Yoshio Tanaka, Isao Kojima, Tsutomu Ikegami, Satoshi Itoh, Satoshi Sekiguchi
National Institute of Advanced Industrial Science and Technology (AIST)

To satisfy both Big Data (BD) processing and Extreme-scale Computing (EC) on a single HPC system, we propose a global address space architecture over a whole data center, which encompasses the distributed memory and the storage. From hardware point of view, advanced optical interconnect technology will fill the gap between inter-node and intra-node memory access bandwidths, which makes the uniform addressing over computation nodes feasible. The architecture and software perspective, however, is much more unclear. We will look at both hardware and software sides below.

On the conventional HPC systems, the inter-node I/O bandwidth is about 1/10 of the intra-node memory access bandwidth. Both of the bandwidths have been increasing over years, though I/O pin-bottleneck is considered to prevent further improvement. The breakthrough for the memory access bandwidth may be brought about by the 2.5D or 3D packaging of processor with embedded memory. Meanwhile, breakthrough on the I/O bandwidth by the optical interconnect technology is much more drastic. Even though the bandwidth of a single optical channel is at most 25~100Gbps, space division multiplexing (i.e., multi-fiber cables) and wavelength division multiplexing (WDM) allow us to bundle some tens of channels without complicated cabling. For future BDEC systems, we think the dense-WDM (DWDM) is most promising: the resulting some Tbps interconnection will fill the gap between inter- and intra-node memory accesses, as far as the bandwidth is concerned.

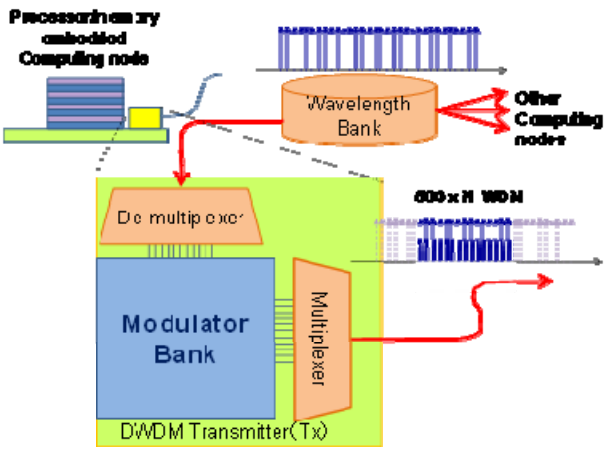

Figure 1 Wavelength Bank (WB)

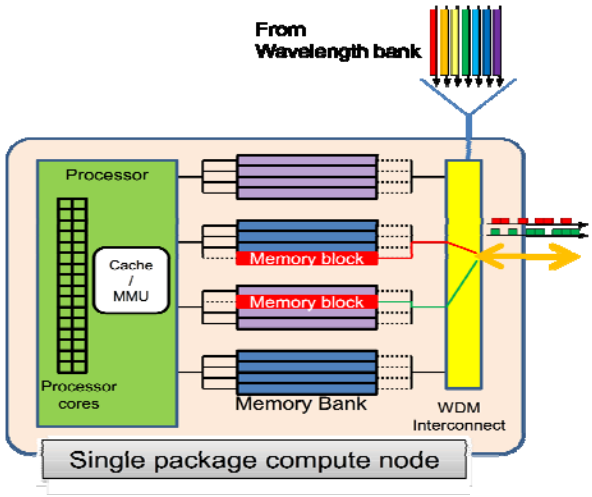

Figure 2 Proposed organization of a computing node

Issues in implementing DWDM in BDEC systems are size, cost and power consumption of DWDM light sources. A light source is required for each wavelength, and to achieve precise wavelength control for DWDM, temperature control and complicated structure is required. Therefore, it is not realistic to implement DWDM light sources in each computing node. To cope with this problem, AIST is proposing a wavelength distribution system using wavelength bank (WB) or optical comb source, and silicon photonics modulator for BDEC systems. WB is a centralized generator of wavelengths for DWDM. The light waves are distributed to computing nodes through

optical amplifiers, and thus no light sources are required for each computing node (Figure 1.). The distributed light is de-multiplexed to each wavelength, modulated, multiplexed again, and transmitted from each computing node. Silicon photonics optical circuits can be used for the whole light wave processing, including modulation, at a computing node. Therefore, size, cost and power consumption can be quite small, and hybrid implementation with electrical circuits is easy. Up to 50Gbps/channel (i.e. wavelength) modulation may be possible by future silicon photonics modulators, and since around 100 channels can be used in a fiber, a total of 5Tbps bandwidth can be realized in a fiber. This bandwidth is more or comparable to memory access bandwidth of a future single package processor-memory node. In addition to using electrical switches which use WB in the same way, such high bandwidth DWDM signals can be switched in one bundle by fiber cross connect switches, or can be switched separately by wavelength selective switches.

To fully utilize the huge I/O bandwidth realized by DWDM, one possible way is to allow direct access to memory by I/O. As shown in Figure 2., main memory is divided into memory blocks, and each memory block can be accessed either from the processor or the I/O at a time. Multiple memory blocks can be sent/received simultaneously using multiple wavelengths. The size of a memory block should be determined considering network paths setup time and data transfer overheads. If the size is 4MB, a block can be transferred in about 1ms at the rate of 50Gbps, and parallel transfer of up to the number of wavelength channels is possible.

There are many issues to realize the BDEC system described above, not only in hardware but also in software. When such structure is employed, software architecture including operating systems, programming models and memory systems should be re-designed too. Management of memory blocks becomes quite important. One possible option may be to map memory blocks to a global virtual address space as shown in Figure 3. Storage can be also mapped to the address space. Memory block transfers can be modeled as a RDMA operations, and for fault resilience, multiple copies of a memory block should be kept in the system. Application programs can access local memory, remote memory and storage class memory through a unified memory interface while considering the difference in the performance. Such system has a strong impact on the structure of an operating system and runtime systems. For example, kernel organization (e.g., hybrid of light-weight and general purpose kernels), data access abstraction on a global virtual address space, fault tolerant/resilience, memory block management scheme, and network resource management (e.g., optimal wavelength scheduling and optical path switching/routing) should be carefully considered. We will conduct a feasible study of the design of both architecture and system software of such system.

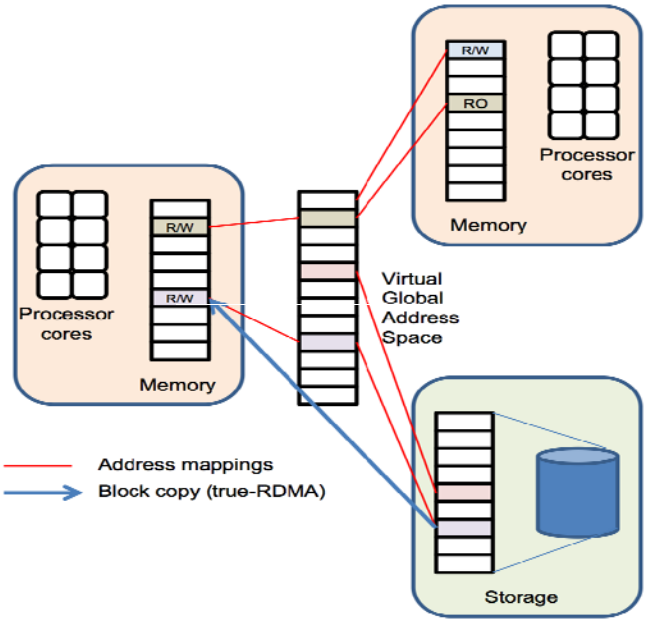

Figure 3 Mapping to global address space