THE 3D TORUS ARCHITECTURE AND

THE EUROTECH APPROACH

a Eurotech HPC Division Short Paper

by

Giovanbattista Mattiussi, Paul Arts, Davide Piccolo
The ability of supercomputers to progressively run jobs faster meets the computational needs of both scientific research and an increasingly higher number of industry sectors.

Processor power is centerpiece to determine the performance of an HPC system, but it is not the only factor. One of the key aspects of parallel computers is the communication network that interconnects the computing nodes. The network is the one guaranteeing fast interactions between CPUs and allowing the processors to cooperate: this is essential to solve complex computational problems in a fast and efficient way.

Together with speed, HPC systems are increasingly asked to be more available. Downtime can affect a high performance machines quite badly. It appears clear that a reliable average machine with a great uptime is better than a high performance one with a low MTBF (mean time before faults): ultimately, the former would process more jobs in a week than the latter.

One additional challenge with large systems is scalability, so the ability to add nodes to a cluster without affecting performance and reliability or affecting them as little as possible. Petascale and then exascale installations require and will require hundreds of thousands of cores to efficiently work together.

It is also paramount for future machines to consume less energy, being cost and availability of electrical power an issue that is becoming the most demanding challenge to overcome along the road to exascale computing.

**3D torus connectivity**

In a computer cluster, the way the nodes are connected together could provide some determinant help to solve the above mentioned issues. Despite being available for quite a while, the torus architecture has now the potential to surge from niche application to mainstream. This is because, like never before, we face nowadays some severe challenges posed by a raising number of nodes. The problem, before being one of performance, is one of topology and scalability. The more a system grows, the more fat tree switched topologies show limits of cost, maintainability, consumption, reliability and, above all, scalability.

Connecting nodes using a 3D Torus configuration means than each node in a cluster is connected to the adjacent ones via short cabling. The signal is routed directly from one node to the other with no need of switches. 3D means that the communication takes places in 6 different "directions": X+, X-, Y+, Y-, Z+, Z-. In practical terms, each node can be connected to 6 other nodes: in this way, the graph of the connections resembles a tri-dimensional matrix.

Such configuration allows the addition of nodes to a system without degrading performance. Each new node is joint as an addition of a grid, linked to it with no extensive cabling or switching. While scaling linearly, with little or no performance loss, is strictly true for those problems that heavily rely on next neighbor communication, it is also true that, avoiding switches, hundreds of nodes can be added without causing problems of clogged links or busy fat tree switchleafs. This comes without considering that the addition of a node in a large system happens with much less working and potential troubles on a 3D torus network than on switched fat tree one.
The pairwise connectivity between nearest neighbor nodes of a 3D Torus configuration helps to reduce latency and the typical bottlenecks of switched networks. Being the connections between nodes short and direct, the latency of the links is very low: this affects positively the machine performance, especially for solving local patterns problems, which can be effectively mapped onto the matrix mentioned above. The switchless nature of the 3D Torus facilitates fast communication between nodes.

Switches are also potential points of failure. Decreasing their number should improve the operational functioning of a system. In other words, the 3D Torus makes a system more agile, so more prompt to react to failures: if a connection or a node fails, the affected communication can be routed in many different directions. The inherent nature of the 3D Torus is the connectivity of each node to its nearest neighbors to form a tridimensional lattice that guarantees multiple ways for a node to reach another one.

Eliminating costly and power-hungry external spine and leaf switches, as well as their accompanying rack chassis and cooling systems, torus architectures also contribute to reduce installation costs and energy consumption.

Applications

When it comes to applications that can fully benefit from the 3D Torus configuration we could touch one of the caveats of this intelligent connection schema.

The maximization of the performance of the 3D Torus takes place with a subset of problems which is rather large but specific. These are local pattern problems, which typically deal with modeling systems whose functioning/reaction depends on adjacent systems. Typical examples are computer simulations of Lattice QCD and fluid-dynamics. More in general, many Monte Carlo simulations and embarrassingly parallel problems can exploit the full performance advantage of the 3D Torus architecture, making the range of possible applications quite vast, especially within the field of scientific research.
Problems that require all to all dialogue between nodes are less prone to exploit the full performance of the 3D torus interconnection. However, independently from the type of application and problem, the 3D torus still bears the massive advantage of scalability and serviceability, contributing also to increase the availability of the systems and reduce power consumption. In case of large systems, it may be so advantageous to resort to the 3D Torus architecture that the perfect match with the problem that better maximize the computational performance may well become secondary.

The Eurotech approach

It is rather interesting to analyze what Eurotech, a leading European computer manufacturer, has done with the 3D Torus network of their supercomputer line Aurora. Eurotech wanted to leverage the 3D Torus benefits for their high end products, but at the same time leave to the users the flexibility and the freedom to run all the applications they need. Taking in account these diverging characteristics, Eurotech and its scientific partners took and approach called Unified Network Architecture in designing the Aurora datacenter clusters. This fundamentally means that the Aurora systems have 3 different networks working in concomitance on the same machine: 2 fast independent networks (Infiniband, 3D Torus) and a multi-level synchronization network.

The coexistence of Infiniband and 3D Torus facilitates flexibility of use: depending on the application, one or the other network can be utilized. The synchronization networks act at different levels, synchronizing the CPUs and de facto reducing or eliminating the OS jitter and hence making the system more scalable.

Torus topologies have traditionally been implemented with proprietary, costly application-specific integrated circuit (ASIC) technology. Eurotech chose to drive the torus with FPGAs, injecting more flexibility in the hardware, and to rely both on a GPL and on a commercial distribution for the 3D torus software. The 3D torus network is managed by a network processor implemented in the FPGAs, which interfaces the system hub through two x8 PCI Express Gen 3 connections, for a total internal bandwidth of 120Gbs.
Each link in the torus architecture is physically implemented by two lines (main and redundant) that can be selected (in software) to configure the machine partitioning (full 3D Torus or one of the many 3D sub-tori available). In this way, redundant channels allow system repartitioning on-the-fly. The possibility of partitioning the system into sub-domains permits to create system partitions that communicate on independent tori, effectively creating different execution domains. In addition, each subdomain can benefit from a dedicated synchronization network.

As an example of partitioning, if a backplane with 16 boards is considered, the available topologies for the partitioning of the machine in smaller sub-machines with periodic boundaries (sub-tori) are:
- Half Unit: 2 x [1:2*NC] x [1:2*NR]
- Unit: 4 x [1:2*NC] x [1:2*NR]
- Double Unit: 8 x [1:2*NC] x [1:2*NR], Rack: 8 x 2*NC x 2, Machine: 8 x 2*NC x 2*NR
- Chassis: 16 x [1:NC] x [1, 2, 2*NR], Rack: 16 x NC x 2, Machine: 16 x NC x 2*NR

Where
- NC : Number of chassis in a rack (8).
- NR : Number of racks in a machine (from 1 to many hundreds)

Partitioning, FPGAs, redundant channels, synchronization networks are some of the unique characteristics that Eurotech wanted in the torus architecture to create Intel based clusters with the flavor of a special machines.