

A New Approach to Clustering

Distributed Federated Switches

1.0 Introduction

Clustering provides a means to build powerful and scalable computing resources using industry standard servers and high performance low latency interconnects. As computer architects have better understood the potential bottlenecks of high performance cluster communications patterns, switching and networking topologies have evolved to allow truly impressive levels of scalability. Several clusters have now achieved 10 Teraflop performance (meaning the cluster is able to calculate ten trillion floating point operations per second)¹. The challenge for system architects is to consider both theoretical issues as well as the realities of the switch system building blocks and synthesize into a practical strategy to develop clusters with the best possible price/performance trade-off. This has meant that the design space available to cluster architects has been necessarily constrained by the switching infrastructure available for high performance interconnects. The current state of the art strategy to assemble large scale clusters using “Federated Switches” reflect these constraints. As always the assumptions leading to the current “optimal” strategy must be continuously analyzed, and if new technologies change these underlying assumptions, then new, possibly different, strategies should be employed to achieve an optimal solution. InfiniBand is precisely such a new technology that provides fundamentally new capabilities, thereby perturbing the status quo, and suggesting new strategies to achieve the next level of optimized high performance compute clustering.

2.0 Federated Switch - The Status Quo

Currently the state of the art in high performance compute clustering suggests a strategy using a “Federated Switch” to achieve large scale computing clusters. A “Federated Switch” (shown in Figure 1, “Federated Switch Diagram,” on page 2) has been defined as *“a packaging solution, which enables very large networks to be implemented with two stages of switch chassis. Although the switch is now physically distributed between multiple chassis, this partitioning is not visible to applications, as the basic switch network topology is unchanged.”*

1. The 1100 Node Virginia Tech InfiniBand cluster recently became only the third cluster to reach this key milestone, achieving 10.3 Teraflops performance.

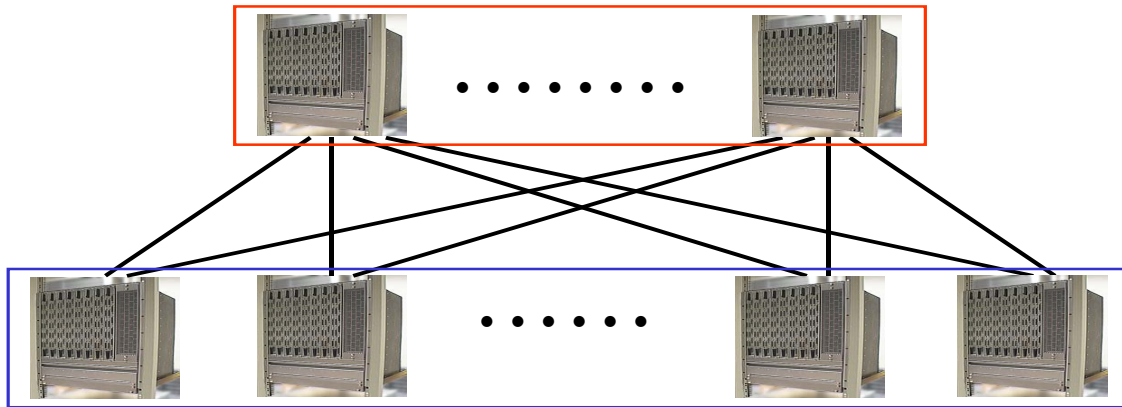


Figure 1. Federated Switch Diagram

This seemingly simple definition of a “Federated Switch”, in fact, considerably over-simplifies the details and underlying assumptions leading to this strategy for assembling large scale clusters. The definition seems to imply that large scale clusters can be assembled from two-stage networks. In fact, what is not made clear here, is that the “switch chassis” themselves are implemented as multi-level networks of lower level crossbar switch elements. Indeed, in the typical case, these switch chassis are implemented as three-stage networks. Thus, in fact the “two-stage” network may, in fact, actually require twelve hops to connect between two end nodes. These additional network stages equate directly to a greater number of silicon devices, increased latency, increased power, and increased cost.

In reality, the traditional Federated Switch approach suffers from a number of disadvantages:

- Large, highly centralized switch chassis do not allow for distributed switching within the rack
- Increases the average and worst case number of hops between nodes
- Poor communications locality
- All end nodes to connect through large centralized switch chassis creates cabling hot spot
- Not possible to uniformly co-locate switches and servers in the same rack
- Increased cable lengths and sub-floor cabling
- Increased potential for Non Uniform Traffic Scheduling (NUTS) Congestion
- Switch port up/down asymmetry complicates correct fabric connectivity

In reality, the federated switch diagram ignores the practical issue of how the servers themselves are connected to the centralized switch. In fact, centralized federated switches make the connectivity issue problematic by increasing cabling lengths and creating a single choke point where all cables converge. This creates very real problems for cable management, cooling, and cluster maintenance and debugging.



Figure 2. Centralized federated switch cabling for 750 node cluster

3.0 InfiniBand - New Technology, New Options

InfiniBand has emerged as the ideal standard for clustering by virtue of delivering 10Gb/sec throughput, low latency, minimal CPU utilization, and offering the first open, industry standard high performance interconnect providing a choice of vendor solutions. InfiniBand is at the forefront of the evolution of clustering from an esoteric technology confined to the worlds most powerful supercomputers to a mainstream data center and high performance computing technology. As such, it enjoys both the economies of scale afforded to mainstream technologies as well as the demands for performance, quality, reliability, and affordability required by enterprise data center customers. While seemingly these demands conflict with the over-arching requirement of performance, in fact, it has become clear that in the long run mainstream technologies actually achieve higher levels of performance than the so called focused, proprietary high performance solutions. The evolution of the CPU market provides a good example of this, where volume shipments of x86 based CPUs have driven capital and R&D investments allowing these processors to outstrip the performance of other custom processors such as the Alpha CPU.

Similarly, InfiniBand is benefitting from widespread adoption in data center applications and leveraging the resulting economies of scale to take advantage of the latest process technologies. Using advanced 0.13um processing technologies both reduces power and increases the number of transistors that can be integrated on a single device. This in turn allows the development of switch chips with unprecedented levels of performance and port density. As will be shown, the increased port densities of a single chip crossbar switch device leads to some interesting new alternatives in the design of large scale clusters. In fact it is the combined market opportunity of both the data center and high performance computing applications that justifies the investment in these advanced processing technologies, and has allowed InfiniBand to leapfrog other technologies.

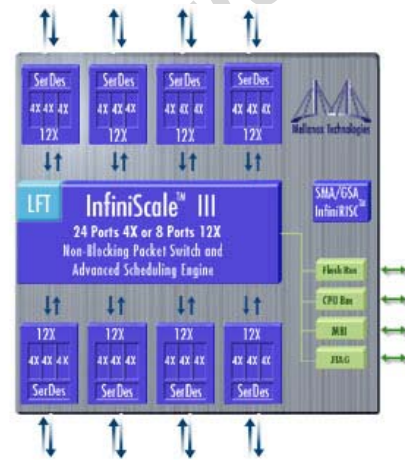


Figure 3. Non-Blocking InfiniBand Crossbar Switch with 24 full wire speed 10Gb/sec ports

4.0 Distributed Federated Switch (DF Switch) - A New Approach to Large Scale Clusters

A Distribute Federated Switch (DF Switch) extends the concept of a conventional federated switch by further distributing the fat tree topology, thereby enabling switch crossbar elements to be as close as possible to server nodes. The DF Switch approach distinguishes between level1 and core switches at a fundamental level (see Figure 4 on page 3). Level 1 (leaf) switches consist of a single highly integrated crossbar chip, packaged in a compact 1U chassis, while only core switches are packaged in larger 9-12U chassis enclosures. Both Level 1 and Level 2 switches are entirely symmetric, with no distinction between upstream and downstream ports, thus greatly simplifying fabric connectivity and setup.

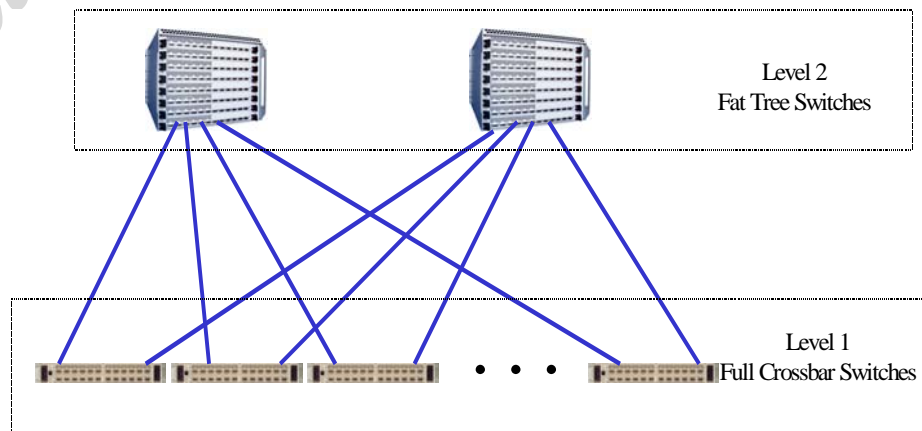


Figure 4. Distributed Federated Switch with Fat Tree L2 and Cross Bar L1 Switches

Furthermore, the DF Switch provides the key advantage that the compact nature of the 1U Level 1 switches allows them to be distributed throughout the cluster. In addition to the Level 1 switch itself, other fabric infrastructure (cabling, management, etc.) can also be distributed and packaged locally with servers - allowing a “switched server” rack. This “switched server” rack unit can be pre-assembled and wired within a rack and delivered as a turn-key, fully functional cluster sub-system - ready to be used as-is to build even larger clusters. This pre-packaged unit can take advantage of *a priori* knowledge of cable lengths, and thus, does not require cable length overhead to account for unknown rack to rack cable routing requirement.

The DF Switch approach results in the following key advantages:

- Fully symmetric configurations for both level 1 and level 2 switches (no notion of upstream or downstream)
- Fewer number of average and worst case hops
- Better cluster communications locality
- Good matching between level 1 switch port density and chassis server capacity
- Eliminates switch tautology (redundant redundancy)
- More efficient silicon utilization
- Less Expensive
- Cabling: Simpler, Less Costly, Easier to Manage
- Simplified Cluster Deployment Logistics

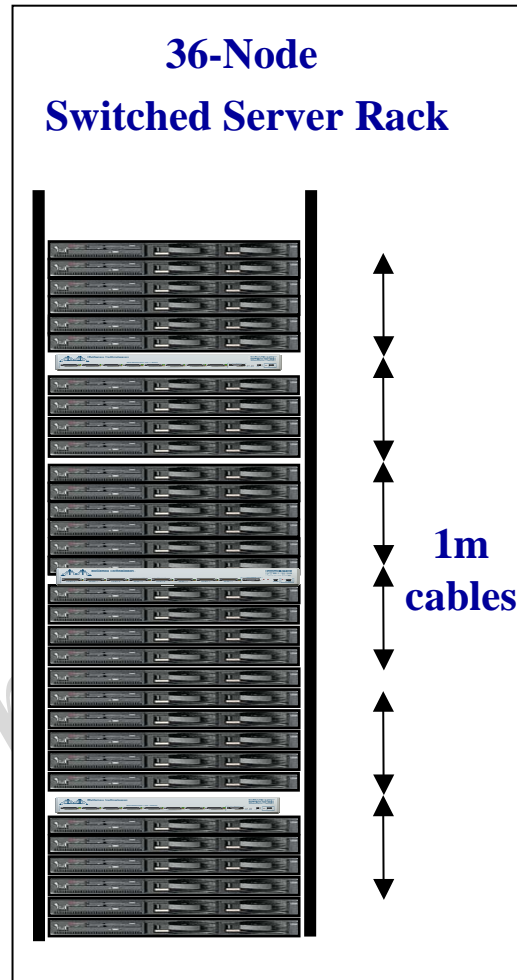


Figure 5. Switched Server Rack Unit

4.1 Why hasn't this been done before?

The key requirement for building DF Switches is a high port density single-chip crossbar. A crossbar port count of 20-24 ports is the minimum to make this approach feasible and economical. Crossbar building blocks with fewer ports require multi-chip implementations of the 1U level 1 switches, making such implementations cost prohibitive and thereby imposing the constraints that switches be implemented with much larger chassis in order to amortize the overhead of the additional silicon cost.

By dramatically increasing the number of ports available in a single chip cross bar element, 1U Level 1 switch units can now be developed with 24 ports available. Twelve of these connected to server nodes and the remaining twelve available for inter-level connectivity.



Figure 6. 24 Port InfiniBand Single Chip 1U Switch

Other proprietary switch technologies have not been able to exploit this avenue for developing clusters offered by DF switches simply because the port density of the crossbar building blocks at their disposal, is not adequate to build a cost effective single chip Level 1 switch. For example, a switch crossbar element with only 8 ports does not provide a sufficient number of ports to amortize the cost of the power supply, cooling, and chassis packaging of a 1U switch platform and thus must be combined into multi-chip platforms.

Achieving high port density single chip crossbar switches is, however, not that simple. One of the keys is to have very high performance serial communications I/O, for without this, the pin count becomes prohibitive. For example, InfiniBand has defined a 2.5Gbaud/sec I/O signalling rate requiring only four differential signals in order to achieve a 10Gb/sec link speed. By contrast, Quadrics uses a much lower signalling rate and requires 40 active signal pins in order to achieve a similar link performance.

This seemingly minor difference has a profound impact on the level of switch integration achievable. Figure 7 on page 5 compares the active signal pin count for InfiniBand and Quadrics. Due to the high signalling rates, each active signal must be complemented by an adequate number of passive pins (power, ground, control, etc.). Thus, current packaging technology imposes a practical limit of around

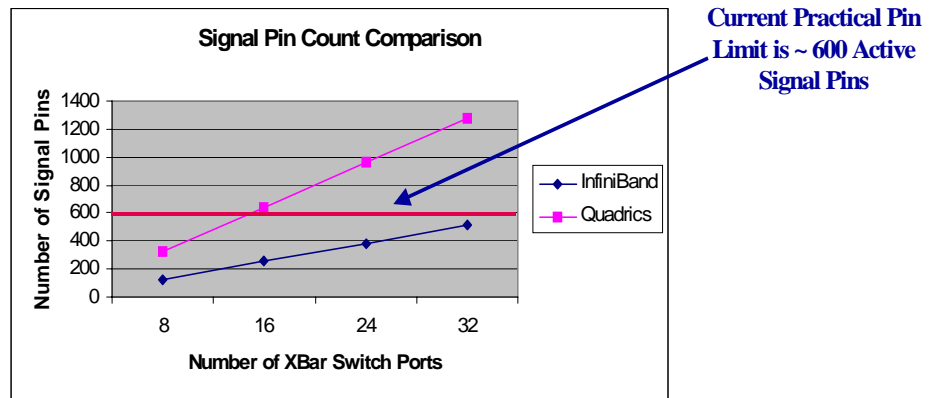


Figure 7. Pin Count vs. Number of Ports Comparison

600 active signal pins for a single integrated switching element. This graph demonstrates how the InfiniBand crossbar switch devices available are offering 24 ports, while for Quadrics, the maximum density is only eight ports. With the announcement by the InfiniBand Trade Association of next generation links speeds of 5Gb/sec and even 10Gb/sec, this port density advantage is expected to continue and grow.

Furthermore, with the DF switch model, the connections between Level 1 and Level 2 switches are completely independent. Thus, for example, link aggregation can be used inter level links, thereby overcoming potential hot spot issues related to non uniform traffic scheduling (NUTS) issues. Furthermore, as they become available, advanced double data rate (5GHb/sec DDR) links can be used for inter-level connections. This enable even more efficient constant bi-sectional bandwidth (CBB) topologies to be built. Basically, 16 single data rate connections between Level 1 switches and host servers and 8 double data rate connections between Level 1 and Level 2 switches provide fully matched bandwidth and greatly decrease the required number of switches to build a given cluster size.

5.0 Summary

Federated Switches have provided the best mechanism to build large scale high performance computing clusters, however, suffer from serious drawbacks related to latency, cabling, congestion, poor locality, and logistics challenges. Advances in InfiniBand technology and semiconductor processing have resulted in the ability to produce crossbar switch elements with much higher port densities. These new high port count switch devices enable a new type of Distributed Federated (DF) Switch to be built which provides significant advantages over the traditional Federated Switch. These advantages include: reduced latency, better locality, reduced chip count, lower cost, simplified logistics, etc. Taken as a whole, these advances in InfiniBand technology and product suggest the Distributed Federated switch as the best approach to building large scale high performance computing clusters.