



## Enabling Efficient Performance at Scale: QLogic IFS 6.0

Addison Snell

May 2010

*White paper*

### EXECUTIVE SUMMARY

A persistent need for increased performance, subject to ubiquitous budgetary constraints, has helped to establish Infiniband as the preferred system interconnect for high performance computing (HPC) clusters. But the same market dynamics that enthroned commodity clusters has also shifted the burden of scalability away from system vendors and onto end users and application programmers. The pursuit of efficient performance at scale is now a software battle.

With InfiniBand Fabric Suite 6.0, QLogic is seeking to differentiate itself from other cluster interconnect vendors with a collection of software features that aim to improve organizational productivity. The components of IFS 6.0 each play independent roles in the optimization or administration of a high-performance cluster. The features of IFS 6.0 fall into three categories:

- **Virtual fabrics**, which allow administrator to establish up to 16 classes of service, so that higher-priority jobs do not get held up by lower-priority system traffic
- **Intelligent switches**, which use adaptive and dispersive routing techniques to reduce and avoid network congestion
- **Support for multiple topologies and vendor-specific MPI libraries**, allowing users to take advantage of both their own environment-specific optimizations and the benefits of IFS 6.0.

Individually these tech-heavy features are interesting to networking wonks who like to post high computational efficiency scores. In combination they represent a suite of a productivity tools that are potentially of interest to the C-level executive who is driven to optimize the use of all resources – computational, capital, and human – to enable new organizational insights.

### MARKET DYNAMICS

In the HPC industry, the dominant purchase criterion for new systems is consistently price/performance – or more appropriately, performance/price, with buying organizations considering which solutions will allow them to get the most real work done within their existing budgets. “Budgets” is intentionally plural in this context, because there are multiple budgets to be considered: not only the capital budget, or what it costs to acquire the system, but also an operating budget to support the system, a facilities budget to power and house the system, and an administrative budget to use the system. Ultimately there is a question of which solution can enable the greatest engineering or scientific insights within these constraints.

This unrelenting drive to maximize performance over multiple cost variables drove the industry through multiple technology transitions, such as from vector to scalar architectures in the 1980s and from RISC-based, symmetric multi-

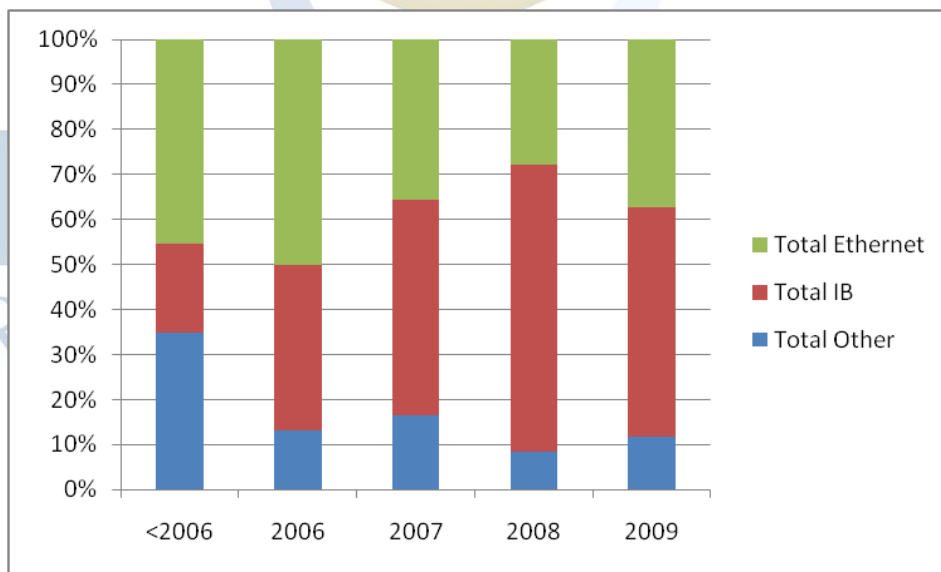
processing (SMP) to x86 clusters in the late 1990s and early 2000s. The dominant share of today's HPC systems are built on commodity hardware and software, but rather than slaking the industry's thirst for greater efficiency, the current reality has left many users striving to find how they can achieve differentiated performance based on standard hardware.

### The Emergence of Infiniband as a Performance Standard

As clusters became the standard HPC architecture, one fundamental performance/price driver was the system interconnect selection, with options segregating into one of two camps: the Ethernets, who found acceptable levels of performance with a lower-cost, IT standard, and the Proprietaries, who saw value in paying a premium – a premium not only in acquisition but also in administration and support – for the bandwidth or latency characteristics of Myrinet, Quadrics, Dolphin, or another performance-focused specialty. The decision was a classic tradeoff of performance vs. cost.

In the mid-2000s, Infiniband began to change the economics of that decision, offering a cluster system interconnect option that for many applications provided a superior performance profile but which was a multi-vendor open standard that was easier to support in an IT environment. Infiniband quickly drove a wave of consolidation in supplanting the majority of proprietary system interconnects in HPC for any organization that saw value in selecting something other than Ethernet for a high-performance system. InterSect360 Research studies have shown that the majority of system interconnects installed since 2007 have used Infiniband as a system interconnect. InterSect360 Research expects this trend to continue in the near term, with minor surges in Ethernet (as in 2009) or Infiniband (as in 2008) driven by generational improvements.

Shares of Primary System Interconnect Types among Surveyed HPC Systems, by Year Installed



Source: InterSect360 Research HPC Site Census Data, 2010

The adoption of standards across the HPC industry has left system and component providers with the competitive challenge of differentiating their offerings without departing from the standard. This challenge extends to both Ethernet and Infiniband providers, who bear the yoke of scalability as the creators of the architectural backbone of the cluster.

### The Burden of Scalability

The drive for efficient performance at scale is an increasing burden. The transition from shared-memory SMPs to distributed-memory clusters increased maximum possible performance per dollar but shifted the burden of scalability from system to programmer. Scalable applications were re-coded into message-passing interface (MPI) models that broke problems into pieces that discrete nodes could work on cooperatively. The necessary inter-node communication for these MPI applications – part data exchange, part communications overhead – is the traffic traveling over the system interconnect switches.

Today there is another technology transition that takes the challenge to yet another level. As microprocessors move into the multi-core phase of development, they introduce another level of parallelism into a cluster. Now there are not only multiple nodes on a switch and multiple sockets (or processors) in a node, but also multiple cores on a socket. For many applications, this can lead to even more inter-node communications as programmers strive to keep cores fed by close-in data.

One constraint of scaling MPI applications is that they necessarily run at the speed of the slowest node, because inter-node communications need to be synchronized. Application programmers therefore may go to great lengths to try to keep each node, processor, and core working, by refining application meshes, pre-fetching data, or streaming multiple instructions whenever possible. Server and software vendors release their own MPI methodologies that are meant to provide optimal performance for their own operating environments.

Delays in inter-node communications can have a devastating effect on performance, and HPC-using organizations also look to different architectural configurations for their larger clusters. Fat-tree, hyper-torus, and mesh topologies represent different methodologies for arranging switches and nodes into a communication hierarchy. Different architectural topologies may be better suited to different workflows, especially as the scale of an application increases.

Yet this is a recipe that can be easily overcooked. At some point, too much administrative cost is going into the optimization of the system, either by application programmers going to great pains to channel traffic effectively or by IT personnel constantly monitoring and manually rerouting traffic. There are diminishing returns to the human capital that is spent on the problem.

For interconnect vendors, and for Infiniband vendors in particular, these end-user challenges create the opportunity for differentiation. Infiniband is an interconnect for those organizations who are seeking differentiated performance for their switches, and there is an opportunity to provide features that extend beyond the qualities of the fabric itself to build intelligent switches that can assist in the quest for efficient performance at scale.

## IFS 6.0 FROM QLOGIC

Seeking to differentiate itself from other Infiniband providers, QLogic has created a suite of software designed to maximize real productivity at scale for a wide range of high-throughput workloads. Each of the software components addresses its own piece of the optimization spectrum, but QLogic markets them collectively as InfiniBand Fabric Suite (IFS). The newest version, IFS 6.0, aims to assemble industry-leading intelligent features for network management and optimization for high-performance workloads at scale, with minimal intervention from system administrators, with the goal of improving efficiency of both people and hardware, ultimately leading to greater organizational productivity.

IFS 6.0 is designed to address the challenges faced by those seeking efficient performance for their applications, regardless of system vendor, topology, or workload. By automating features for system optimization and management, QLogic creates a scenario for maximizing performance without overtaxing operations, administration, or support. The target result is greater efficiency in the drive for scientific or engineering insights.

The features of IFS 6.0 fall into three main categories: virtual fabrics, intelligent routing, and support for multiple MPI libraries and architectural topologies. Each of these plays a specific role in optimizing for efficient performance at scale. Collectively, this set of features gives organizations the ability to make the network choices that will best serve their particular workflows, while remaining confident that they can still optimize their applications' performance. Ultimately IFS 6.0 gives organizations the possibility of greater application efficiency and performance while simultaneously reducing the burden of system administration.

#### *Virtual Fabrics: Classes of Service*

One of the challenges in achieving scalable MPI performance is that critical computational messages can get stuck in network traffic behind less urgent messages. A rush-hour traffic jam is a good analogy. Consider the vehicles that might be all using the same road: a father taking a child to a play date, a mother trying to get frozen food home from the grocery store, one person trying to make it to the movies on time, another trying to catch a plane, someone who is late for work, and someone who is in the back of an ambulance on the way to the emergency room. Everyone might want to get to where they're going, but nevertheless they have different levels of urgency.

With virtual fabrics, the administrator can designate up to 16 distinct "classes of service," different priority levels for different types of network traffic. When network congestion is encountered, the computational message passing for the current engineering job can be given higher priority than the data loading for a project that isn't scheduled to begin yet. The effect is like turning on the ambulance's siren so that other traffic allows it through. Ultimately all the network traffic gets to where it is going, but it is processed in an order that is optimal for overall operational efficiency. The highest-priority messages get processed with the least delay, keeping critical applications from being slowed by network overhead.

#### *Intelligent Routing: Adaptive Routing and Dispersive Routing*

In many cases the better way to deal with a traffic jam is to take a detour. With many network topologies for systems with multiple switches, there are redundant pathways that connect common points. With adaptive routing, intelligent switches see network congestion and divert traffic around it.

Consider an inner-city street grid in which you wish to go two blocks northeast. There are many paths that are roughly equivalent, and some of them (such as north-east-north-east and east-east-north-north) use completely independent paths and intersections. IFS 6.0 enables intelligent switches to monitor the traffic around them so that they can send messages on uncongested paths.

Because it is a critical enabler for network performance, adaptive routing is not a concept unique to QLogic, although there are different approaches to the problem. Some adaptive routing mechanisms rely on a subnet monitor (sometimes called a subnet manager) to poll the switches for traffic information; the subnet monitor then gives routing instructions to the switches. In QLogic's implementation, the intelligence is built into the switch itself, allowing each switch to make decisions without waiting for instructions and also eliminating the overhead of communicating with the subnet monitor.

Another difference in the IFS 6.0 adaptive routing scheme is that the QLogic switches see data in terms of flows rather than individual packets. In an MPI workflow, there are some discrete packets of data that should be transmitted together and in order, like a parade (albeit a very fast one) or other procession of vehicles on a road. By viewing these complete

flows in making routing decisions, the QLogic implementation helps ensure that data is arriving at its destination in the correct sequence.

Similar to adaptive routing, the goal of dispersive routing is to use the network in the most efficient manner possible. When data bursts do not need to be kept together or processed in a particular sequence, dispersive routing automatically distributes the individual packets over different network paths, effectively load balancing the system. The combination of adaptive and dispersive intelligent routing techniques helps avoid network traffic jams without system administrator intervention.

*Multiple Support Options for Optimization: Vendor-Specific MPI Libraries and Architecture Topologies*

In the quest for efficient performance at scale, hardware vendors, software vendors, and HPC-using organizations test out a myriad of optimizations for their systems. Many vendors release MPI libraries that contain tricks and shortcuts for navigating their particular operating environments. And depending on how data and communications flow for a particular application, different topologies for connecting a hierarchy of switches can provide scalability benefits.

IFS 6.0 is not locked into any particular MPI library or system topology. End users can implement whichever MPI libraries or fabric topologies best suit their application workflows, and they can still take advantage of the class of service and intelligent routing benefits of IFS 6.0.

**QLogic IFS 6.0 Feature/Benefit at a Glance**

IFS 6.0 Feature	What It Does	Productivity Benefit
<b>Virtual Fabrics / Classes of Service</b>	Allows administrators to specify up to 16 different priority levels for jobs	Higher priority jobs are processed without waiting for lower priority traffic
<b>Adaptive Routing</b>	Intelligent switches sense fabric congestion and reroute data flows to avoid it	Avoids network traffic jams before they happen by preventing backups
<b>Dispersive Routing</b>	Packets that can be separated are sent across the network on different paths	Automatically load-balances the system to avoid congestion
<b>Multi-Vendor MPI Libraries</b>	Supports MPI libraries that are specifically optimized for the operating environment	Programmers and applications are free to use any optimizations with IFS 6.0
<b>Multiple Topologies</b>	Supports any architectural topology (fat tree, torus, mesh, etc.) at scale	End users can select the topologies that best suit their workflows

*Benefits of IFS 6.0*

With IFS 6.0, QLogic has recognized that high-performance hardware has become an industry standard, and therefore differentiated efficiency will come from software. Infiniband has made it easier to build scalable, high-throughput clusters, and IFS 6.0 builds intelligence into the system fabric to drive more efficient performance at scale. With the continuous thirst for optimal productivity with limited budgets -- capital, operating, and human -- the benefits of IFS 6.0 target the most persistent challenge in the HPC industry.

## INTERSECT360 RESEARCH ANALYSIS

HPC users have established the role of Infiniband as a high-performance cluster interconnect, and InterSect360 Research expects its presence to continue throughout our technology forecast horizon, tick-tocking share increases with Ethernet based on generational advancements in bandwidth and latency. Among Infiniband vendors, differentiation in performance will shift increasingly to software features that allow end users to optimize for efficient performance with minimal intervention.

There is a clear market opportunity for QLogic with IFS 6.0 in the current market dynamics, and the supporting competitive dynamics will become increasingly important as challenges in parallelism continue to grow through the wholesale transition to multi-core processor architectures. Science, engineering, and analytics firms are exploring new architectures, multiple vendors, and virtualization techniques in attempts to optimize performance without continual intervention.

In the longer run, InterSect360 Research is also monitoring the growth opportunities manifest in the emerging use of HPC technologies in application areas that transcend the traditional HPC boundaries of science, engineering, and analytics. These “Edge HPC” areas have many aspects of traditional HPC workflows but tend to be more IT than R&D. Specific Edge HPC application areas are:

- *Complex event processing*: Scanning multiple, real-time data feeds for patterns that demand an action. Examples: high-frequency trading, airport security modeling, credit card instant fraud warning.
- *Business process optimization*: High-end business intelligence and data mining applications to improve operational efficiency. Examples: inventory tracking and optimization, logistics, buying pattern analysis.
- *Virtual environments*: Creation of artificial, real-time interactive digital worlds. Examples: online games, virtual reality, augmented reality.
- *Ultrascale business computing*: Supercomputers running applications that do not fit other definitions of HPC workflows, but supercomputing is implied strictly by the level of scale. Example: online retail is not an HPC application, but Amazon.com has a supercomputer.

These Edge HPC applications have typically skewed toward Ethernet-based system deployments, but IFS 6.0 has the potential to help make QLogic Infiniband a good fit for this group of users. The solution is standard but high-performance; it conforms to IT standards and uses virtualization and class-of-service techniques to deliver measurable quality-of-service metrics.

With IFS 6.0, QLogic is meeting the challenge of delivering a differentiated solution. The challenge for QLogic will be to communicate adequately how a concoction of obscure and apparently disparate software features – each of which individually is technically deep – can contribute to organizational productivity. Lifting the messaging from a low-level system administration feature to a high-level CTO benefit is difficult, and many managers are not eager to spend an hour discussing fabric optimization.

Nevertheless QLogic has taken a big step in understanding and promoting the importance of software features in making systems more efficient. If QLogic is able to effect a change in discussion from raw performance to overall productivity, then there will be a significant opportunity for the company to find a competitive advantage in IFS 6.0 that could lead to adoption and growth in both traditional and Edge HPC environments.