

The Case for InfiniBand in AdvancedTCA

AdvancedTCA (ATCA) at its core is a fabric agnostic architecture, and one of the fabrics defined early on was InfiniBand via the ATCA 3.2 sub specification. As ATCA systems move towards production it is likely that the majority of these systems are ATCA 3.1 or Ethernet based systems, but there are limitations in this kind of deployment. Others propose Advanced Switching as the “King of All Fabrics”, but truly there may be no king of fabrics in ATCA. Time to market, bandwidth, latency, and even system level costs are some areas in which InfiniBand may have a surprisingly good fit for ATCA.

Some target telecommunication engineers are outright hostile to the usage of ATCA in their company’s application. The chief reason for that hostility remains bandwidth. Engineers report that 1 Gb/s Ethernet was slower than their current proprietary solution and that 10G Ethernet was not the timely solution they needed. On the InfiniBand side, one x4 InfiniBand port can provide 10 Gb/s of bandwidth utilization, which would fully utilize one ATCA channel, providing the timely bandwidth demanded. Other fabrics have plans for these bandwidth levels; ATCA 3.1 Ethernet via Option 9 and ATCA3.4 Advanced Switching via Option 3 can match this in terms of raw performance. The difference is the timing of the solution. With InfiniBand, the full bandwidth available on a simple FR4 backplane is realized today, without limitations on node boards or switches. Both Ethernet and Advanced Switching solution suffer from the timing aspect on availability of proper ATCA centric node and switch solutions.

Another key factor that affects performance is the time it takes a single packet to leave the host and reach the destination. This factor is known as packet latency. InfiniBand offers its high bandwidth with low latency. Packet construction time is an area where InfiniBand exceeds thereby reducing latency. The InfiniBand host channel adapter (HCA) does this directly, without host support and does not carry the overhead of Ethernet, where a complex software protocol stack is necessary to insure reliable delivery. A well-designed packet switch is required to achieve low latency as well. InfiniBand answers this with a switch that can cut through a packet from input port to output port by looking at the first 8 bytes of the header, before the packet has even fully arrived. Other switch architectures will buffer the entire packet before routing, resulting in increased latency. This cut through routing does mean forwarding a bad packet but that doesn’t cause a fabric or application failure because the packet will be checked at its destination.

One area where there is much discussion occurs in the form of cost of the fabric. Interesting shell games occur when this subject is discussed. Ethernet does have a clear message about per port hardware cost being lower than other fabrics. But per port charges are only partial solutions therefore only a partial story of the cost. 1G Ethernets need multiple ports to even approach half the bandwidth of InfiniBand. Alternatively, some fabrics rely on future technology that is not available in the scale and port counts needed for ATCA. Where this future cost is not known - it can’t be extrapolated nor inferred from “sister” technologies. A real world example based on existing ATCA nodes and switch, an apples to apples comparison, of system cost is possible. If a fully loaded InfiniBand ATCA3.2 Option 1 (two switches and 12 CPU nodes) system is compared to the exact same configuration of a loaded Ethernet ATCA3.1 Option 4 system, the total system cost of the InfiniBand system is around \$1000 greater than the Ethernet system; around about 2% of the total system cost. This minor increase comes with a 10X increase in fabric bandwidth and similar performance increase in fabric latency. When one looks to the CPU utilization or packet construction off loading, a conservative estimate of 30% of available CPU cycles are used in packet construction in this real world Ethernet system. A greatly more liberal estimate of 10% CPU cycles for packet construction on the InfiniBand side translates into about \$1500 savings from wasted CPU cycles. That saving reoccurs in the form of more power for applications running on the CPUs. InfiniBand allows processors to do what they were purchased to do, and not munging packets. Certainly link aggregation and off-load engines increase the likelihood of the matching performance and limiting CPU usage but these performance enhancements belay the apparent cost per port benefit of Ethernet. Additionally, the silicon cost of InfiniBand switches are admittedly more expensive mainly because of market size differences, but one major difference is InfiniBand management software is open source. The ease and general familiarity customers required on the Ethernet side requires a complicated laundry list of RFC support. Frequently, the management cost of the Ethernet are not included when raw per port hardware costing is used, and this is one reason the system level cost of Ethernet is not realized from its per port lead. The result is “The Great Fabric Shell Game”. Instead of finding the ball under the shell, we are ask to forget cost additions of off load engines, multiple port aggregations,

needed to argue performance limitation and to buy into wild speculation on future enhancements, and to ignore development Ethernet software management cost in effort to show Ethernet is the low cost fabric of choice. Per port cost saving is a partial solution and therefore one must look toward total system cost and the performance gained. Today, InfiniBand is the clear winner when timely system level price for performance is taken into account.

Advanced Switching attempts to point to the economies of scale due to the reuse of the physical and link layer technologies to predict lower cost. While there is value in reusing existing technology, it is clear that these economies will not be achieved. AS will have much higher prices than volume PC based PCIExpress products like motherboard chipsets, and add-in cards. AS devices, will sell in much lower volume and will have much higher performance, reliability and support designed into them and will therefore demand higher cost. AS and PCIExpress are different technologies that will be deployed on different classes of devices. PCIExpress is not at all suited for system level ATCA construction via ATCA redundant links; AS is required. AS means management software, where PCIExpress requires none thereby changing the market, node architecture, chip design and adding cost. PCIExpress is a great technology for desktop systems, and even enhances InfiniBand by allowing reduced BOM costs via memory free HCA solutions. The AS "economies of scale" argument is simply not valid, it is not PCIExpress. Additionally, current port counts on AS switches, mean that some blocking will occur on early deployment of AS on ATCA, so any timely matched performance argument is hurt here as well.

Ultimately, there may be no king of ATCA fabrics. Ethernet, Advanced Switching, InfiniBand, Rapid I/O and others can and likely will all coexist in ATCA. Pitting one fabric against another is fun and the heated discussions that arise are interesting. ATCA is designed to its core as a fabric agnostic architecture. As engineers and technologists, it is our job to ensure that once all the tradeoffs have been studied, the best solution is chosen and for some ATCA applications that is assured to be InfiniBand.

Article Written by Joe McDevitt, VP of Technical Development.

All trademarks and tradenames are the property of their respective owners.

Whitepaper/Editorial Contact Information:

Joel Deer; Marketing Communications; marketing@dtims.com; 601.856.4121

Customer Contact Information:

DTI Sales; Diversified Technology, Inc.; 476 Highland Colony Parkway, Ridgeland, MS 39157
1.800.443.2667; sales@dtims.com; www.dtims.com