White Paper

SYNOPSYS®

Optimizing Storage and I/O for Distributed Processing on Enterprise & High Performance Compute (HPC) Systems for Mask Data Preparation Software (CATS)

Glenn Newell, Sr. IT Solutions Mgr,
Naji Bekhazi, Director of R&D, Mask Data Prep (CATS)
Ray Morgan, Sr. Product Marketing Manager, Mask Data Prep (CATS)

Abstract

Semiconductor, or chip, manufacturing represents one of the most complex manufacturing processes in the world. This process, which includes building miles of thin electrical/signal wires connected to millions of transistor switches (all at sub-nanometer dimensions) would never be achieved if it were not for the 'magic' and power of EDA CAD tools; which include software applications and solutions from system level (product) design through semiconductor manufacturing and are provided by Synopsys; a global leader in semiconductor design, IP and manufacturing solutions.

This paper will focus specifically on Synopsys EDA software tools, process flows, and compute resources used for the "Mask Data Preparation (or MDP)" process flow area of semiconductor manufacturing. (Synopsys product; CATS®). Often referred to as the critical link from design to manufacturing CATS is a software package which converts semiconductor design data into a useful format for making the circuit templates (also referred to as "photomasks") for semiconductor manufacturing. Topics discussed will include:

1) Semiconductor design and scaling trends fueling increases in device complexity and file size growth
2) Design file sizes and challenges to MDP software
3) MDP compute system network challenges and requirements for large file processing
4) Advanced high performance compute systems for MDP scaling
5) Results and Best practices for Enterprise and HPC solutions on CATS
6) How to work with Synopsys to implement these type of solutions

1. Introduction

As chip geometries continue to shrink to the 45nm node and beyond, the resulting increases in design complexity and chip pattern density have fueled a data explosion on advanced semiconductor designs. This extends product development cycles and potentially impacts product yield.

The use of aggressive OPC and resolution enhancement techniques (RET) required for continued semiconductor device scaling has resulted in an exponential increase in the output database file sizes (which is the input to MDP). Estimates of advanced OPC file sizes are expected to surpass one Terabyte (1TB) within 2 years.

This increase in OPC file sizes directly impacts subsequent data processing, such as mask data preparation (MDP), creating increased challenges for integrated device manufacturers, foundries and mask shops to limit the negative impact of increased RET to MDP turn around times (TAT). Controlling this TAT is a critical challenge for the MDP software, compute systems and the supporting network infrastructure.

Mask Data Prep application is typically one of the most I/O intensive processes in the design cycle. One can add more compute power, via distributed processing, which allows for reasonable turnaround time despite the increased design file sizes, but the problems shifts and then exacerbates the load on storage and interconnect (network) – or I/O (Input/Output)

Carefully examining, and then mitigating the I/O and storage bottlenecks that are created with modern Mask Data Prep distributed processing tools can lead to reduced turn around times, on the order of 4x, and a reduction in the amount and cost of infrastructure required.

2. IT Challenges with I/O Intensive Distributed Processing

2.1 Network File System Bottleneck

Even with the advent of multi-core processors, the amount of required RAM makes it impractical to run Mask Data Prep within a single compute server in a reasonable amount of time.

The result is the need for Distributed Processing (DP). As all DP worker processes need to access a common input file, and at some point the DP output must be collected and merged into a single output file, a network vs. direct attached file system is needed

This immediately creates a performance challenge in that network file systems are typically “slower” than locally attached disks. This is in large part due to cluster interconnect (network) bandwidth and latency. GigE network bandwidth is smaller, and the latency higher, than Fiber Channel or SCSI.

2.2 Simultaneous Read from a Single File

A second performance challenge arises in that all the DP processes are simultaneously reading from the same input file. With increasing numbers of DP processes a CPU or backend bottleneck in the Network File Server (NFS server) is created.

Figures 1 and 2 illustrate these limitations during a CATS DP run using a single GigE attached Network File Server. During the MDP Fracture step, the multiple workers reading the exploded data file simultaneously overtax the file server, and begin to saturate the single GigE interconnect to the server. While the theoretical Bandwidth of GigE is ~ 125 Megabytes per second, Fig.2 shows that less than 90 Megabytes per second is the maximum achievable throughput under these conditions.

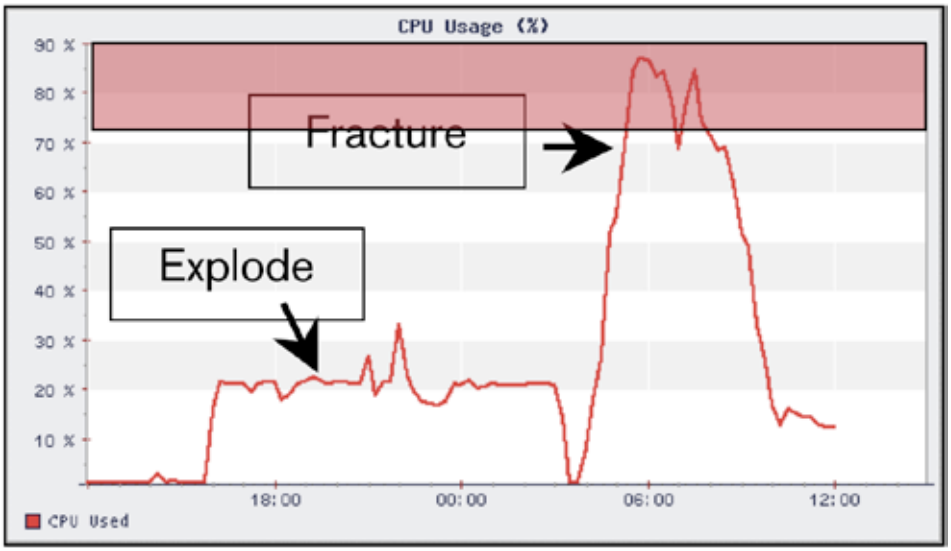


Figure 1: Large CPU Count Fractures overtax NFS Server CPU.

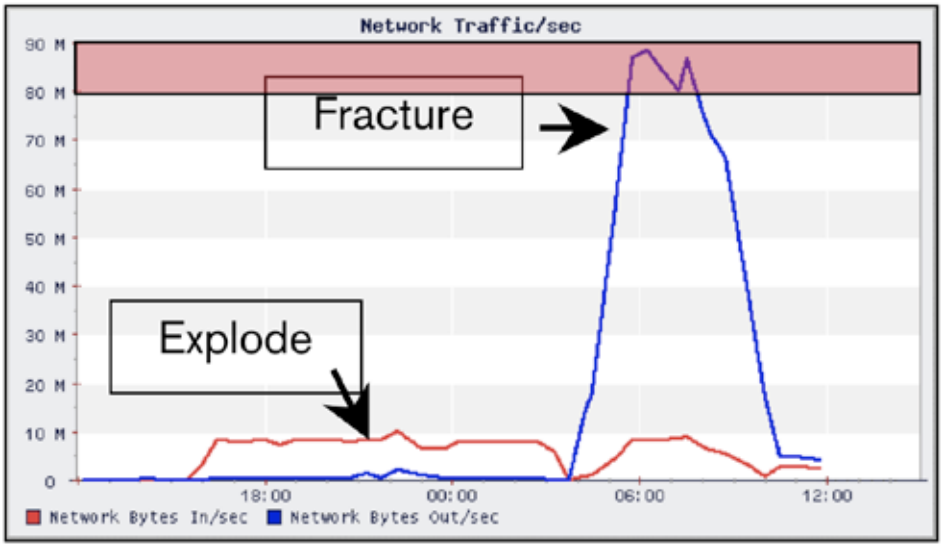


Figure 2: Maximum Read Bandwidth for Single GigE attached NFS server is ~ 90 MB/sec.

2.3 Network Blocking Factor

The third performance challenge is the blocking factors found in a typical enterprise network between the DP workers (computer servers) and the storage server.

In a typical enterprise network (shown in Figure 3), compute servers are connected to an access layer switch with either GigE or Fast Ethernet. The switch then has a 1 GigE uplink to a building switch, and then a similar setup back down to the network storage servers.

By way of example, given 16 compute servers in a cluster on a common network switch, and a storage server on a separate Ethernet switch, the minimum blocking factor will be 16:1. Only one of the 16 compute servers can talk to the storage server (at full "wire" speed) at a time.

Many large Ethernet switches have internal blocking factors, with each group of eight ports sharing an ASIC with 1 gigabit/second access to the back plane.

Figure 3 illustrates a typical three tiered (Access, Distribution, Core) enterprise network and some of the bottlenecks found in a typical enterprise data center. The following is a summary of the bottlenecks illustrated.

1) Computer Servers connected by Fast, rather than Gig Ethernet
2) Access Layer switches with internal blocking factors
3) Blocking factors created by multiple Ethernet switches, where the uplink bandwidth is less than the sum of the aggregate port bandwidth
4) Network file servers connected with less bandwidth than the aggregate sum of the compute server bandwidths

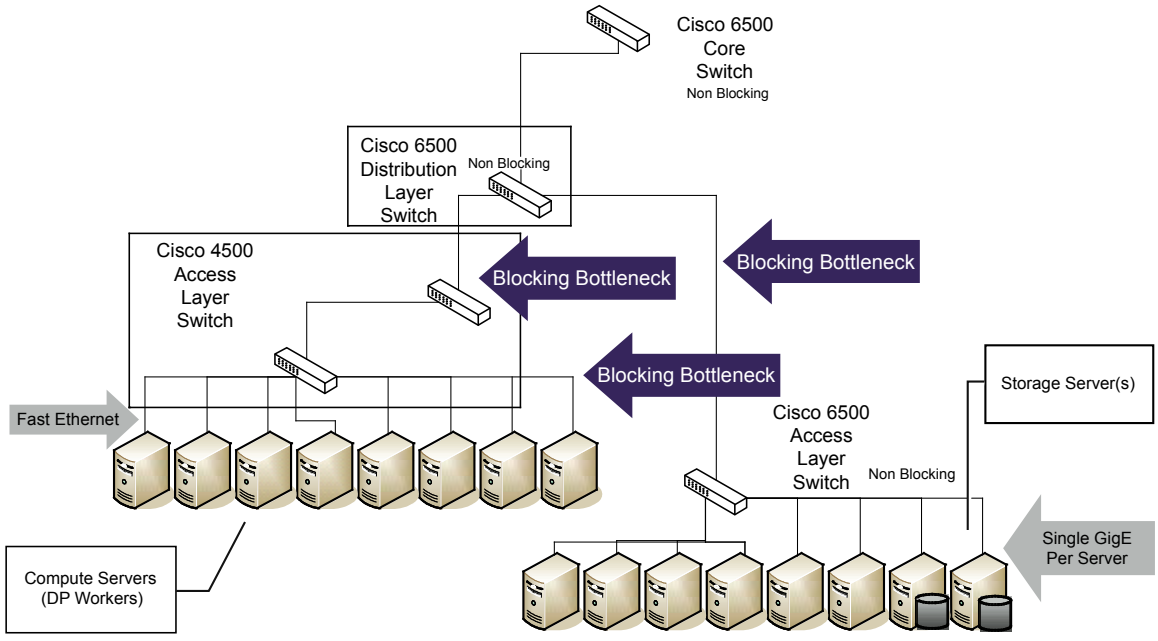


Figure 3: Typical Enterprise Network showing Bottlenecks.

All of these bottlenecks and blocking factors result in DP workers that cannot talk simultaneously, and at wire speed, to the storage server. This can be graphically illustrated with parallel IO tests such as the Pallas Parallel benchmark.

Figure 4 compares a cluster of blade centers with their built in switch (14:1 blocking factor) to compute servers with individual connections to a single Non Blocking GigE switch. With 14 nodes connected to a single uplink the bandwidth falls off as soon as the number of processes (1 process per compute node) goes above 14.

With a Non-blocking single Cisco 6513 GigE switch, bandwidth doesn't fall off even out to 240 processes. This also holds true for small inexpensive 24 port non-blocking GigE switches such as those from Linksys, Netgear, and Dlink, out to the limit of their port count.

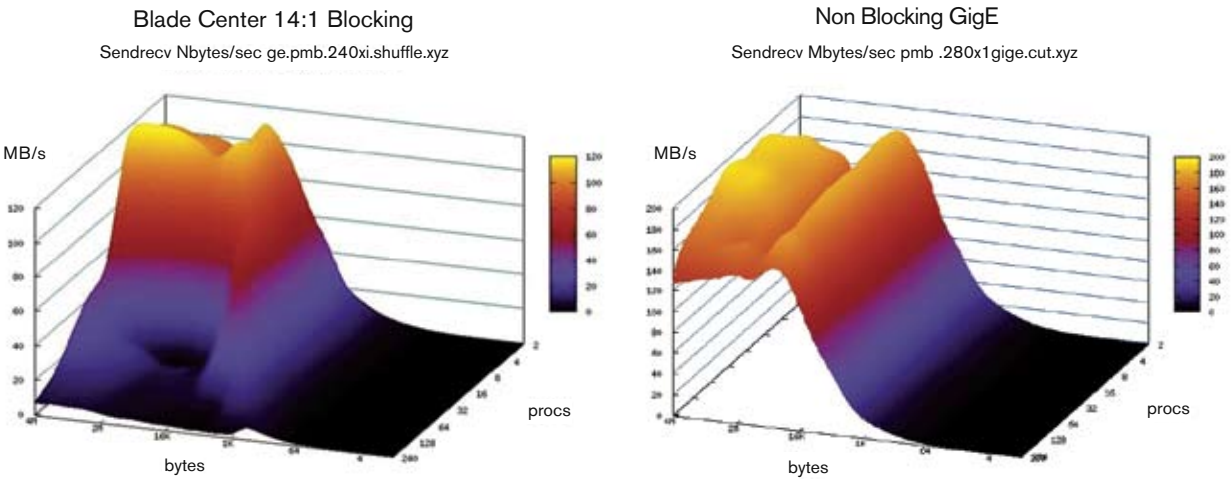


Figure 4: Pallas Benchmark Results for blocking and Non-blocking Network.

3. Recommendations for CATS DP on Enterprise Class Network File Systems

"High Performance" Network Storage Systems and interconnects will be discussed in the in the next section.

For "Enterprise" class systems, the following recommendations can be used in order to mitigate the above performance issues Figure 5 illustrates this architecture.

1. Use a dedicated Enterprise class NFS appliance, such as Network Appliance FAS 3020 or higher
2. Make as many network connections to the NFS appliance as it allows
3. Use a non-blocking, single GigE switch to connect DP Masters, workers, and storage server together.

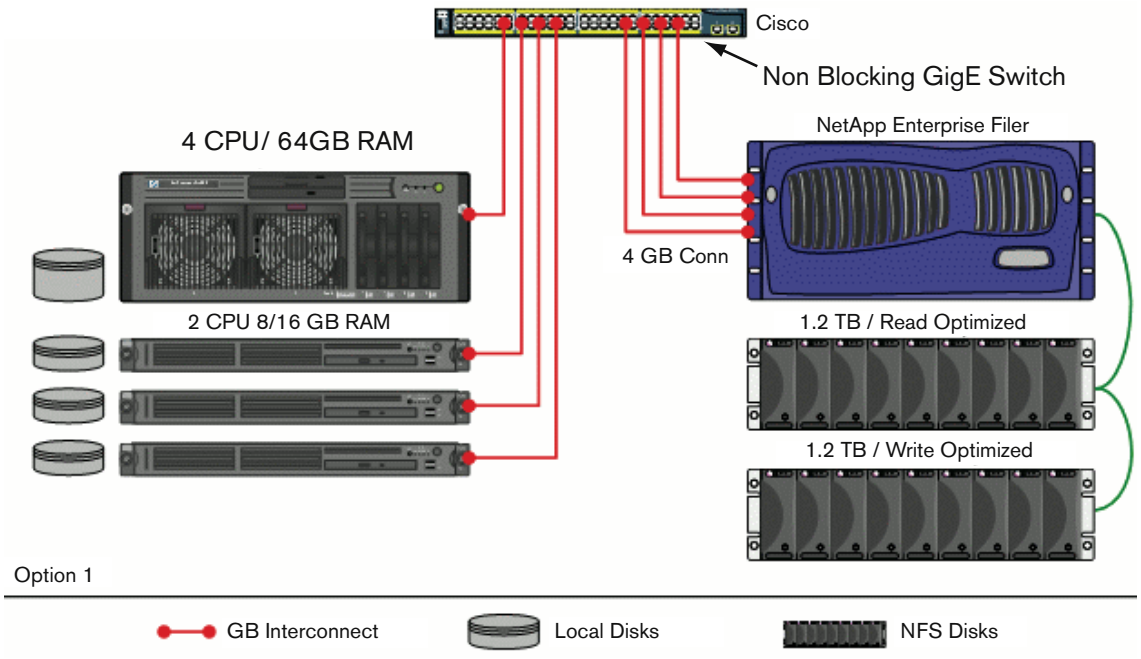


Figure 5: Recommended Architecture NFS + GigE.

The above is the “best in class” architecture for running MDP via distributed processing, given a GigE network and NFS storage. However, this is by no means the limits of currently available IT compute infrastructure. Using the above as a baseline, or a performance of “1”, it is possible to achieve a further 4x increase in performance.

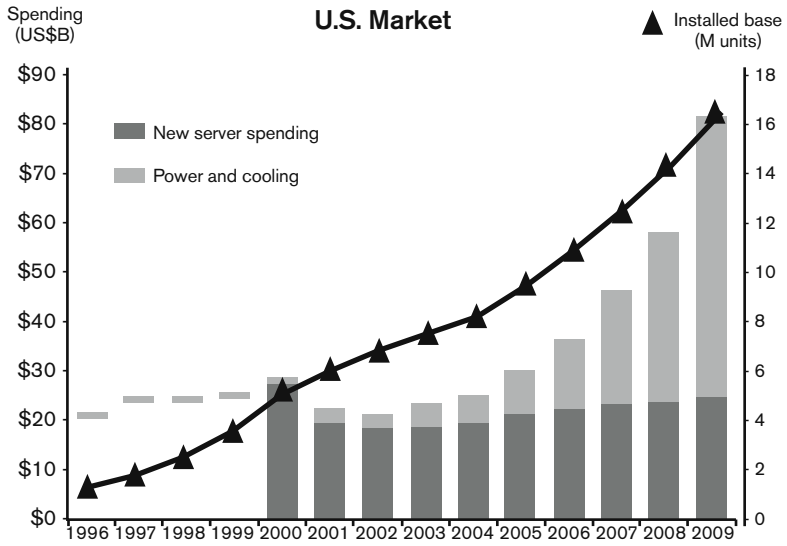
4. Using High Performance Storage and interconnect to achieve 4x MDP performance

4.1 Multicore

The performance of individual CPUs has stopped increasing, due to various thermal, memory, and architectural limits. Instead CPU manufacturers are delivering performance increases via multiple CPU cores inside one package. This implies further scale out in Distributed Processing, further exacerbating DP’s I/O challenges, as well as concentrating more required I/O bandwidth within a single compute server. E.G. if a single core dual CPU compute server can saturate a single GigE connection, then a dual core dual CPU compute server will require more bandwidth.

4.2 Data Center Costs

Beginning in 2007, the Data Center Costs to house, power, and cool compute servers is exceeding their capitol acquisition cost. Therefore it is desirable to minimize these costs by using High Performance Compute technology to minimize the number of compute servers needed to complete a job in the required amount of time. The projected trend of data center costs is shown in figure 6.



Source IDC: 2006, Document #201722, “The Impact of Power and Cooling on Data Center Infrastucture”, John Humphreys, Jed Scaramella.

Figure 6: Design Center Costs: Server, Power & Cooling Trends.

4.3 Design Size

Due to smaller resolution and aggressive Reticle Enhancement Technology (one type often discussed – OPC or Optical Proximity Correction), mask pattern data file sizes continue to trend upward. 100 GB files are already appearing in production MDP flows, and projections for the next few years predict that 1 TB files will not be unusual.

The International Technology Roadmap for Semiconductor or “ITRS Roadmap’ has been predicting exponential growth in file size, although the forecast has been revised downward over time. Practical experience with the CATS product shows the trend in file size increase to be linear, rather than exponential, and entering the 1 TB range. This is illustrated in Figure 7.

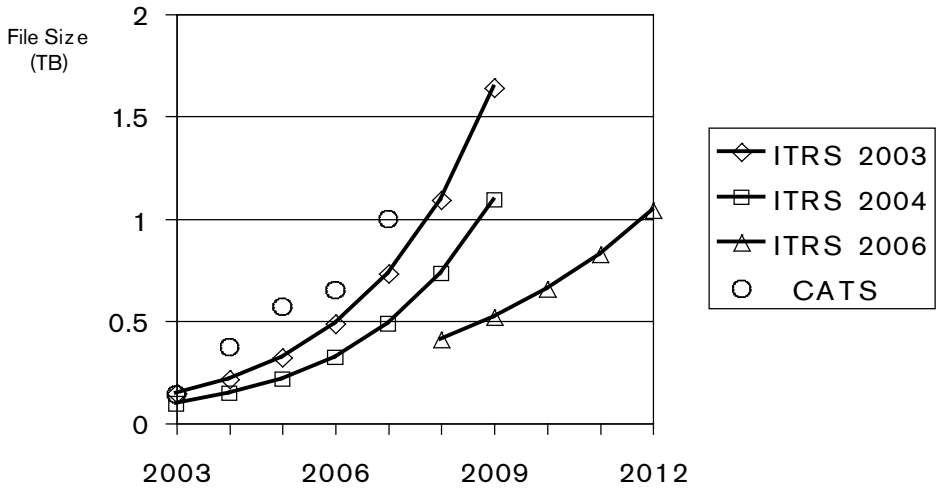


Figure 7: MDP File Size Trend – ITRS & CATS.

4.4 Fixed Turn Around Time (TAT)

Regardless of increasing design complexity and the accompanying file size, the ability of mask shops to "turn" designs in the same amount of time drives their competitiveness. The following table represents estimates from the ITRS Roadmap of both file sizes for OPC data processing (input file for MDP/CATS) and the processing time required for both OPC and MDP. It should be noted here that the times listed are general guidelines. Leading mask shops will use more demanding criteria – typically under 12 hours for MDP processing time.

Category	**2005**	**2007**	**2009**	**2011**
DRAM _ Pitch	80nm	65nm	50nm	40nm
OPC File Size	260GB	413GB	655GB	1040GB
OPC Prep Ime	3-6 days			
MDP Prep Time	1-2 days			

4.5 High Performance Solutions and Interconnects

There are three factors that affect performance over a given network interconnect. These include 1) Bandwidth, 2) Latency, and 32 CPU Overhead. Figure 8 lists the performance factors for currently available interconnects.

Interconnect	FastE	GigE	10GigE	Myrinet	Infiniband 4X SDR
Bandwidth	100 Mb	1 Gb	10 Gb	2 Gb	10 Gb
Latency	1.2ms	60us	10us	8us	4us
CPU overhead	80%	80%	80%	6%	3%

Figure 8: Interconnect Performance Factors.

Of these interconnects, Infiniband has the best or equal performance in all three categories. Bandwidth equal to 10GigE (for Single Data Rate 4x Infiniband, faster versions of Infiniband are now becoming available), the lowest latency, and the lowest CPU overhead (for those applications that use native Infiniband, vs. tcp/ip.).

While the "per port" cost of Infiniband Host adapters, cables, and switches may be higher than some of the other interconnects, the resulting performance offsets these costs by achieving equivalent turn around times with fewer compute nodes.

4.6 Lustre Parallel File System

Parallel file systems allow for the “striping” or spreading of files across multiple file servers. Note that this is in contrast to “clustered” file servers, where each server serves its own file system; parallel file systems serve the same files from multiple file servers, distributing the load.

Lustre has an advantage over some other parallel file systems in that it is a native infiniband application. No context switches are needed for the CATS application to talk to Lustre storage via infiniband. This means that Lustre takes full advantage of the low latency and low CPU overhead of infiniband.

Figure 9 illustrates that a single Lustre server, via infiniband, was able to perform at a rate of 250 Megabytes/second, in the same situation were an NFS server, via GigE, was limited to 90 Megabytes per second.

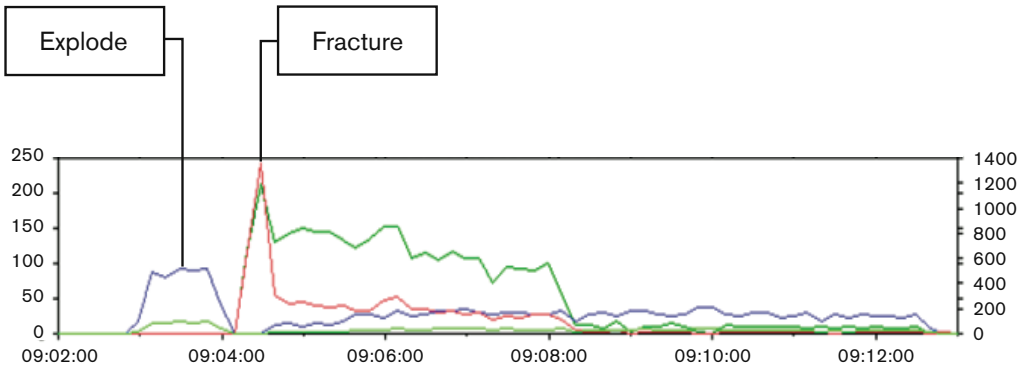


Figure 9: Lustre single server performance.

The advantages of Lustre and infiniband over GigE and NFS are further illustrated in figures 10 and 11. Figure 10 represents the results from an early prototype cluster using Lustre and infiniband where Figure 11 represents the latest series of tests on CATS to determine scalability performance through 116 total CPUs (near linear with superior scalability).

Going back to Figure 10, the results were ploted as the time to complete a test case against increasing numbers of worker CPUs. The right hand y axis shows the time to complete the test case. The top plot in purple is the results for the Lustre infiniband cluster, out to the maximum number CPUs in the cluster, 64. The bottom blue line plots the results for the same type of compute servers, in a GigE NFS cluster.

As the number of CPUs increase beyond sixteen, the performance of the GigE NFS cluster begins to fall off, and actually becomes worse at sixty four CPUs than at sixteen. In contrast, the Lustre infiniband performance continues to increase to the maximum number of available CPUs. The left hand y axis and the yellow line plot the percent improvement of Lustre + infiniband over NFS + GigE ending at 153%, or 2.53x improvement.

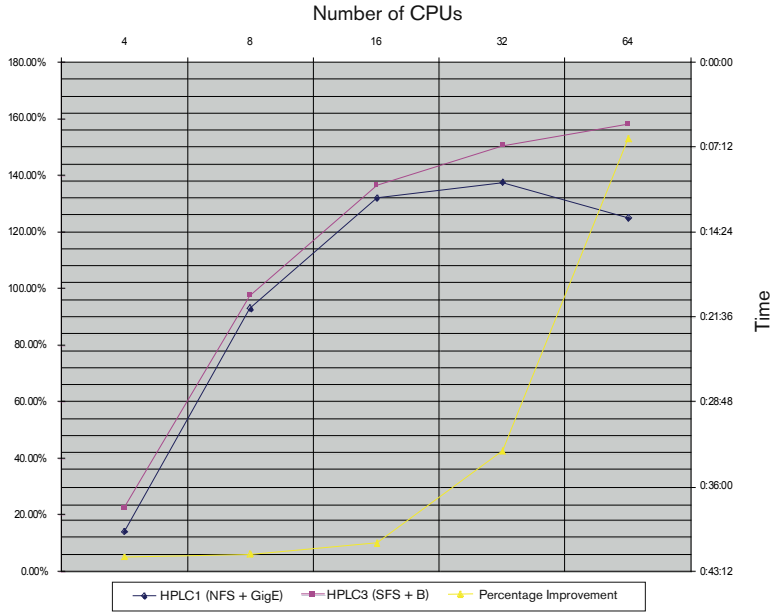


Figure 10: NFS/GigE vs Infiniband/Lustre.

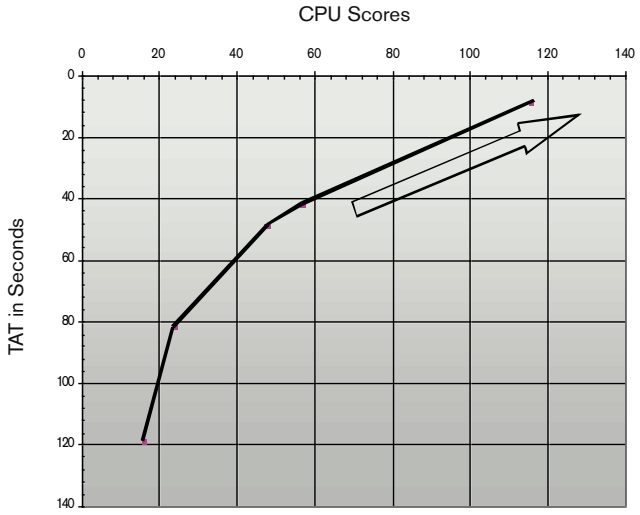


Figure 11: NFS/GigE vs Infiniband/Lustre.

In summary, superior scalability was achieved for CATS through 116 CPUs using a high performance compute clusters. The performance tests demonstrated the enhanced performance for scalability that can be achieved when using Infiniband/Lustre vs NFS and GigE. It should be noted however, that the NFS and GigE enterprise will also continue to develop new enhancements to drive performance gains.

5. The Final HPC Configuration/Solution

For the final solution the storage system was designed to deliver the same amount of per CATS worker process storage performance as seen in the prototype cluster, but scaled to 256 worker processes.

- 17x the storage performance of the GigE + NFS cluster
- 6x the performance of the prototype Infiniband + Lustre cluster
- 8 Lustre Object Storage Servers (vs. 3)
- 160 FCAL spindles 16 TB usable (vs. 72 spindles 8TB)
- Fault tolerant high performance storage array with 8x2 RAID stripe (can survive and perform through multiple disk failures) vs. software raid
- 264 port Infiniband switch with 6 GigE uplinks (vs. 24 IB ports)

5.1 Super Computer Class Storage Performance – Enterprise Price

In order to ensure the storage performance would scale, a "best in class" high performance storage array was chosen, based on it being deployed in 7 of the TOP 10 super computers (at that time), many of which use the Lustre file system. A key point that should be highlighted is in the cost per compute power. When one analyzes the per terabyte cost of this HPC storage versus a normal Enterprise NFS solution it is lower.

Below are illustrations of the HPC solution including Figure 12 which shows a schematic representation of the final configuration and Figure 13 which shows an architectural view (show here for comparison to Figure 5 - the Enterprise class solution)

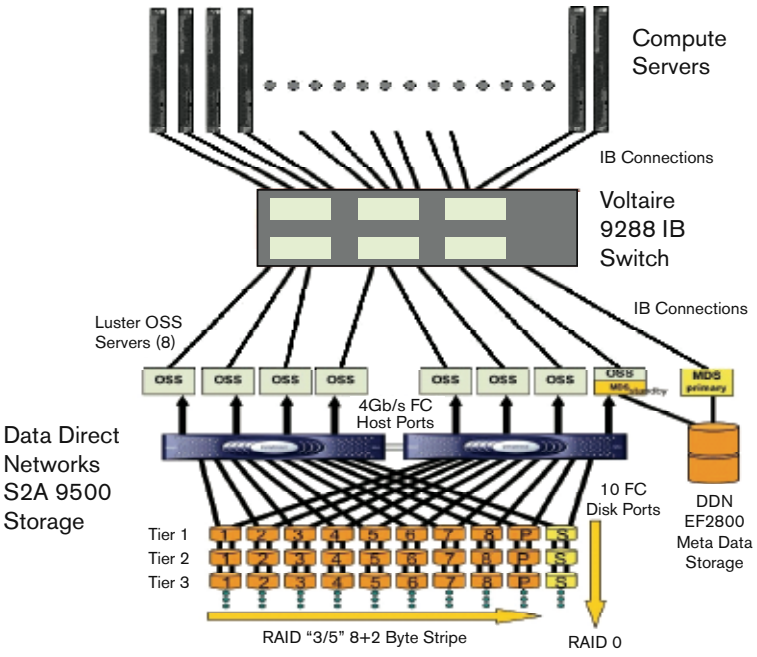


Figure 12: HPC Solution Schematic Illustration.

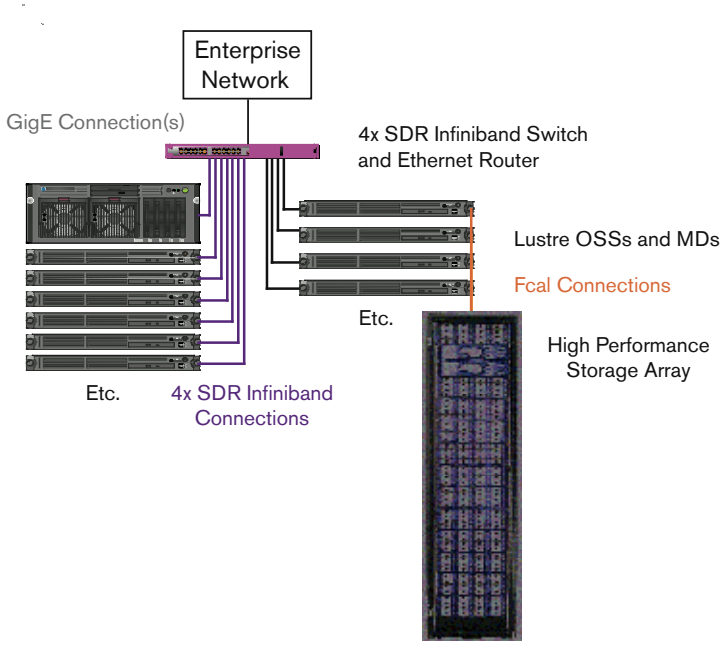


Figure 13: Final HPC Solution Architecture.

6. CATS MDP Results using the HPC Architecture

6.1 Speedup and Scalability

Figure 14 illustrates test run results on three data sets, comparing Lustre + Infiniband to GigE + NFS. Twenty CPU CATS parallel distributed processing was used on both clusters. The Data sets consisted of 300GB GDS files. Figure 15 represents the scalability performance (first discussed in Section 4.6). The speedup and scalability factors would be expected to increase with an increase in the numbers of CPUs

CATS Speedup: 2.75x to 4X improvements (IB + Lustre vs Gig+NFS)

CATS Scalability: Superior: Achieved near linear scalability through 116 CPU's (max tested)

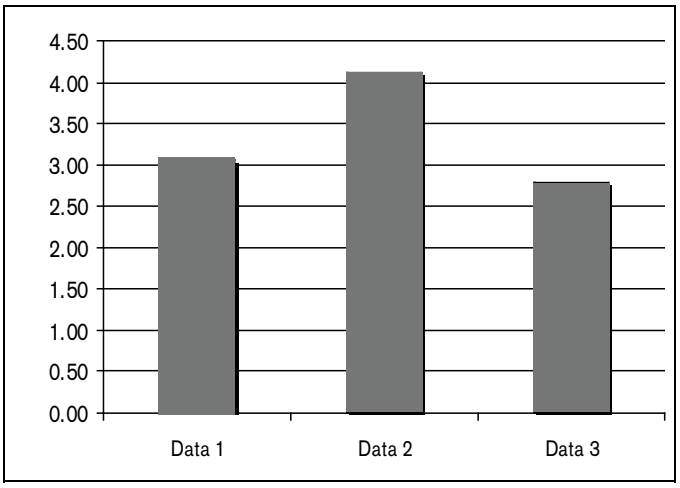


Figure 14: CATS Speedup on HPC.

.

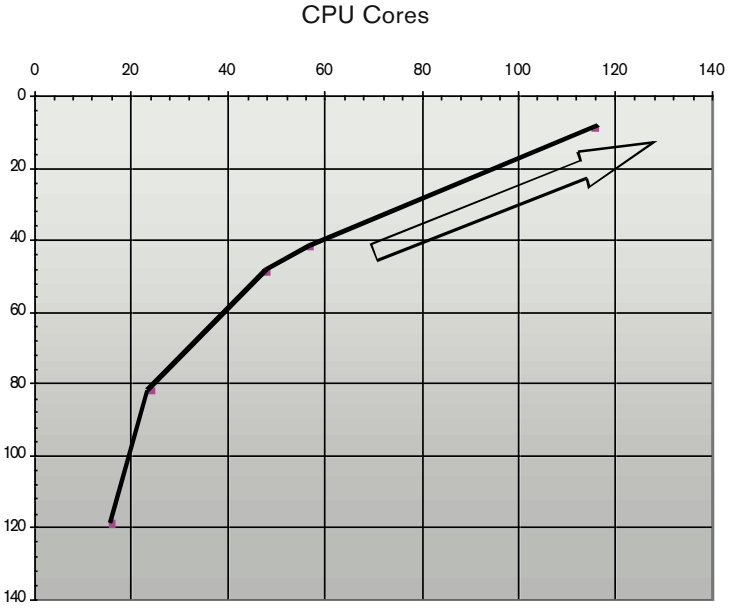


Figure 15: CATS Scalability on HPC.

6.2 CATS HPC and MDP Best Practices

In addition to the HPC system and its performance advantages, one should also be aware of how to leverage the interoperation of both CATS and the HPC such as what Linux OS to use, what parallel file system to use, what type of non-blocking interconnects, etc. Here is a good starter list. For more details one should contact your Synopsys/CATS support person or account manager.

Best Practices:

- Use a Linux based 64bit X86 cluster
- Use a parallel file system such as Lustre
- Use a high speed low latency non-blocking interconnect such as 4x Infiniband
- Use a high performance storage back end such as Data Direct Networks with lots of FCAL spindles
- Scale storage backend, storage servers, and interconnect so that the storage system can deliver an optimized MB/second a minimum of 15 - 25MB/sec (B for Bytes) per compute node (e.g. for 200 compute nodes storage and interconnect should be able to deliver 3 – 6 GB/sec)

Note: 1) GigE bandwidth max = 125 MB/sec, 2) 10 GigE/Infiniband 4x max = 1.25 GB/sec, 3)Infiniband has lower latency (than 10GigE) and other advantages such as the ability of Lustre to use native Infiniband protocols, bypassing the Linux Kernel and TCP/IP stack.

7. Summary and Conclusions

As with most industries, delivering product to market faster and more cost effective is a constant driving influence. With short product cycles and ever increasing product complexities, the Semiconductor Industry pushes these demands to the extreme. Finding best in class solutions to deliver to these demands is what Synopsys is all about.

The HPC solution presented in this paper for the CATS mask data preparation product from Synopsys is world class and delivers superior performance on both speedup and scalability – a critical need for MDP users in the semiconductor manufacturing ecosystem. The HPC solution represents is the results of highly coordinated efforts of multiple teams including Synopsys IT Solutions, CATS Product Teams, and various hardware/software suppliers including Voltaire, Data Direct Networks, Mellanox Technologies, and Cluster File Systems.

Finally, as validation to our expertise and capabilities for designing, building, setup and testing of Enterprise/HPC systems integrated and optimized with software applications such as CATS one should consider the following;

1. Synopsys has designed, built, and deployed multiple HPC solutions for both our in-house needs and for our customers
2. Synopsys was the first EDA software provider in 2006 to be listed as one of the Top 500 Supercomputing companies in the world.
3. Synopsys has built a world class IT Solutions workforce that specializes in building Enterprise and HPC solutions

8. Contact Synopsys

We hope this information presented within the framework of this white paper proved valuable. As it may be true that not all MDP customers would require such an HPC system today, one should use this document as a basic planning document for both your current and future needs. Terabyte file sizes are coming – and the sooner we all prepare the better. For more information go to http://www.synopsys.com/products/solutions/dfm or call 508-263-8006.

SYNOPSYS®

700 East Middlefield Road, Mountain View, CA 94043 T 650 584 5000 www.synopsys.com