Reference Architecture: Multi-Node HPC Workload Scaling with Cornelis CN5000 Omni-Path on Lenovo ThinkSystem Servers

Technical Showcase of CN5000 Omni-Path Operation on Lenovo ThinkSystem Servers

Showcase of real world HPC workloads from a variety of sectors

Tunings needed to achieve best performance for HPC out of CN5000 Omni-Path Network

Studies with public workloads with well understood performance behaviour

Kevin Dean
Conor Elrick

LENOVO PRESS

Table of Contents

Abstract

High-performance computing (HPC) workloads such as computational fluid dynamics (CFD), computer-aided engineering (CAE), molecular dynamics, weather modeling, and materials science increasingly rely on scalable, low-latency interconnects to efficiently utilize growing CPU core counts across multiple compute nodes. As workloads scale, application performance becomes increasingly dependent on the ability of the network fabric to deliver high bandwidth, low latency, and predictable communication performance.

This Reference Architecture (RA) demonstrates how Lenovo ThinkSystem servers, Lenovo Neptune liquid cooling, and Cornelis CN5000 Omni-Path networking can be combined into a scalable HPC platform. The reference design features a Lenovo N1380 Neptune Direct Water Cooled Chassis populated with 8 compute trays totalling 16 nodes of Lenovo ThinkSystem SC750 V4 servers powered by Intel Xeon 6 Processors connected with high speed internode communication handled by Cornelis CN5000 Omni-Path SuperNICs and Switching.

To illustrate expected real-world behavior, we evaluate a diverse collection of publicly available HPC applications spanning CAE, CFD, molecular dynamics, weather forecasting, and computational chemistry. The study highlights both workloads that scale nearly linearly across multiple nodes and workloads whose performance becomes increasingly communication-bound at larger scale. For each workload, we identify recommended network tuning settings and runtime configurations that help maximize application performance on the CN5000 Omni-Path fabric. The results provide practical guidance for system architects, administrators, and technical decision makers seeking to design, deploy, and tune scalable HPC environments on Lenovo infrastructure.

This RA provides both a validated deployment blueprint and application scaling guidance for organizations implementing CN5000-based HPC clusters.

1 Introduction

This paper presents a validated Reference Architecture (RA) for scalable HPC workloads built on Lenovo ThinkSystem servers and Cornelis CN5000 Omni-Path Technology networking. The architecture was implemented and tested in the Lenovo Benchmark Centre where optimised software toolchains and runtime environments were deployed to evaluate the performance and scaling characteristics of representative HPC applications. Using a diverse set of real-world workloads, we examine application behaviour from a single node through a 16-node deployment and identify configuration and tuning practices that maximize performance at scale.

The paper is structured as follows:

- Section 2 describes the hardware architecture used in the reference design, including Lenovo part numbers for building upon the reference setup
- Section 3 discusses the software environment and system tuning recommendations to achieve optimal application performance at scale on ThinkSystem SC750 V4 servers connected with Cornelis CN5000 Omni-path networking.
- Section 4 presents scaling performance for a diverse set of representative HPC workloads, highlighting application behaviour across multiple nodes and identifying configuration settings that deliver the best performance at scale.

This paper follows on from a previously published technical whitepaper named “Installation and Best Practices for Implementing Cornelis CN5000 Omni-Path on Lenovo ThinkSystem Servers” [1] available on Lenovo Press at the link: https://lenovopress.lenovo.com/lp2474-installation-and-best-practices-cornelis-cn5000-omni-path. In this previous publication, we showcased the deployment of a CN5000 Omni-Path RA in the Lenovo Benchmark Centre and performed low-level performance testing. This paper extends that work by evaluating the scaling behaviour of representative HPC workloads in addition to the full architecture design.

2 Hardware Overview

This section provides an overview of the hardware components and system configuration used in the reference architecture.

2.1 Reference Architecture Overview

The reference architecture as depicted in Figure 1 includes:

- A single Lenovo ThinkSystem N1380 Neptune Enclosure
  - 8 Trays / 16 Nodes of ThinkSystem SC750 V4 Servers. Each node contains:
    - Dual Socket Intel Xeon 6980P 128c, 500W Processors
    - 24 (12 per socket) Units of 64GB DDR5 RDIMM Memory operating at 6400MT/s
    - ThinkSystem Cornelis CN5000 Omni-Path 1-Port QSFP112 HFI SuperNIC
  - CN5000 Omni Path 48-port Edge Switch
    - Sixteen 2m links of 400G QSFP112 DAC Passive Cable, one to each compute node
  - 1 GbE Management Switch
    - 3m Cat 5E connection to the CN5000 Omni Path 48-port Edge Switch
    - 3m Cat 5E connection to the N1380 Neptune Enclosure

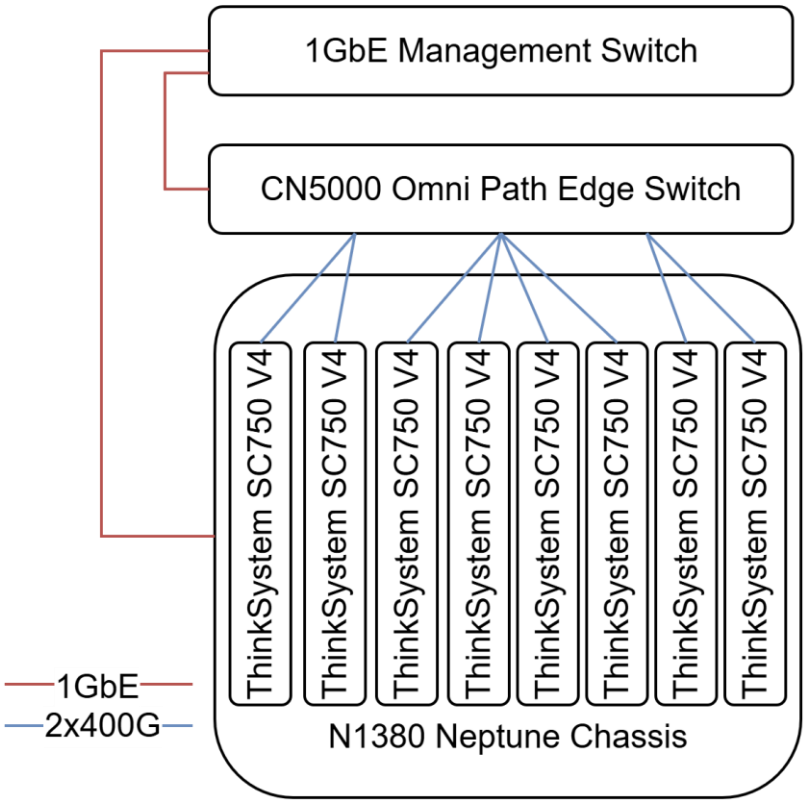

Figure 1: Diagram of Hardware Used in RA Overview

Each of the 16 compute nodes has a 400G link to the Omni Path Edge Switch. The firmware used on the CN5000 Edge

Switch matched EveryScale release 26A[1] where FEC is enabled on all switch ports. Therefore, the sixteen QSFP112 DAC will have the same available features on any of the 48 switch ports and can be connected anywhere. The compute servers all have XCC and UEFI firmware matching the same Lenovo EveryScale best recipe release.

2.2 Bill of Materials

Table 1 presents the Bill of Materials (BOM) for this reference architecture. In this BOM, the main building blocks are given in the rows in bold font, with the rows that follow including components and option choices.

- The first block of the BOM in shows the 16 nodes of SC750 V4 inside the N1380 Neptune Enclosure including the 400G links to the CN5000 Switch
- The two lines which follow are for the CN5000 Edge Switch itself
- The final 3 lines are for the 1GbE management switch and the 2 connections to enable management access to the N1380 Enclosure and CN5000 Edge Switch

Part Number	Product Description	Qty per System	Total Qty
7DDHCTOLWW	**Lenovo ThinkSystem N1380 Neptune Enclosure**		**1**
7DDJCTOLWW	**ThinkSystem SC750 V4 Neptune Tray - 3-Year Base Warranty**		**8**
C2WQ	Intel Xeon 6980P 128C 500W 2.0GHz Processor	4	32
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	48	384
BYK4	ThinkSystem SC750 V4 Neptune Tray	1	8
CB7C	ThinkSystem Cornelis CN5000 Omni-Path 1-Port QSFP112 HFI SuperNIC DWC (Generic FW)	2	16
CBPP	Cornelis 2m CN5000 400G QSFP112 DAC Passive Cable-Straight	2	16
7DMQCTO2WW	**Cornelis CN5000 48-Port OPA 400Gbps Air-Cooled Switch oPSE**		**1**
CC44	Cornelis CN5000 48-Port OPA 400Gbps Air-Cooled Switch oPSE	1	1
7D5FCTOFWW	**NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)**		**1**
BPC7	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	1	1
3798	3m Green Cat5e Cable		2

Table 1: BOM for the Reference Architecture

To scale this configuration beyond a single N1380 enclosure, the remaining 32 ports on the CN5000 Edge Switch can be used to connect additional nodes and the available ports on the 1GbE Management switch can provide additional

[1] Available from https://support.lenovo.com/us/en/solutions/HT518494

management network connectivity. Large deployments beyond three enclosures will require a multi-tier fabric architecture using additional CN5000 switching infrastructure.

3 Toolchain Modules and Performance Tuning

To build and run HPC applications on the RA setup, a software environment containing the required compilers, libraries, and runtime dependencies must first be established. In this section, we describe the creation of these toolchains and the environment tunings needed to achieve optimal performance for HPC workloads. These configurations provide the foundation for the application scaling results presented in the following sections.

3.1 Toolchains and Modules

Before testing application performance on the fabric, it is necessary to create MPI toolchains that provide the compilers, libraries, runtime environment, and supporting dependencies required to build and execute the benchmarks on the system. To provide flexibility for different user environments, we create two MPI-based toolchain modules: one based on Intel One-API[2] and a second based on OpenMPI[3]. Both are configured according to the recommendation outlined in the CN5000 Performance Tuning Guide [2].

For Intel MPI, we download the offline installer from the official webpage and run it to install a shared installation on our system.

```
./intel-oneapi-hpc-toolkit-2025.2.1.44_offline.sh -a -s --eula=accept\
      --install-dir=/hpc/xproj/cn5000_opa/software/intel-oneapi --instance=cn5k-hpc
```

This will create an installation of OneAPI which we can source with the created `/env/vars.sh` script. At the same time as sourcing OneAPI, there are a few environment variables that needs to be set globally for all users on the system. These are:

- `I_MPI_OFI_LIBRARY_INTERNAL=0` to disable the libfabric that comes with OneAPI
- `I_MPI_FABRICS=shm:ofi` to set OFI as the fabric used
- `FI_PROVIDER=opx` to use the OPX provider

For convenience we can set this up as a module file or as a shell script which can be sourced. The Lua module used for our setup is included in Appendix A1 of the CN5000 Setup on ThinkSystem guide [1].

For OpenMPI, after we have obtained a release tar bundle from the official webpage, we built it with a GCC installed already on our system via EasyBuild[4] (GCC v15.2.0).

```
INSTALL=/hpc/xproj/cn5000_opa/software/OMPI_5.0.10
module load eb-env
module load GCC
tar -xvzf openmpi-5.0.10.tar.gz
cd openmpi-5.0.10
./configure --prefix=$INSTALL --with-ofi=/usr --enable-orterun-prefix-by-default
```

[2] Intel One-API publically available from https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html

[3] OpenMPI publically available from https://www.open-mpi.org/

[4] Easybuild available at https://easybuild.io/

```
make -j 128
make install
```

A module file for sourcing the OpenMPI installation and configuring the required environment variables described above is provided again in Appendix A1 of the CN5000 Setup on ThinkSystem guide [1].

3.2 Settings and Tuneables

This section highlights three key configuration settings for achieving optimal performance on the CN5000 Omni-path fabric. Where additional workload-specific tuning is required, those recommendations are detailed alongside the application results in the following section.

3.2.1 Context Sharing

With the OPX setup, there are a maximum of 208 HFI contexts available per node when running multi-node MPI applications. This restricts the user to using 208 MPI ranks per node, unless context sharing is enabled. In newer OPX releases this has now been enabled by default, but in our setup which matches the (at the time) current EveryScale Best Recipe, we need to enable this in our environment for our node with 256 physical cores with the following:

- `FI_OPX_CONTEXT_SHARING=1` to enable the use of shared contexts [Required for our choice of CPU SKU]
- `FI_OPX_ENDPOINTS_PER_HFI_CONTEXT=2` to set 2 MPI end points per context [Sufficient for our choice of SKU]

If the best run time options for a workload requires using less than 208 MPI ranks per node, context sharing can be left in disabled mode.

3.2.2 Bulk Transfer Service

Bulk transfer service (BTS) is a technology that reserves HFI contexts which are used on top of the user allocated contexts in order to improve the communication bandwidth for some workloads. To enable BTS, we need to reserve some of the MPI contexts at boot time which will no longer be available to the user. Within Lenovo Confluent, we can do this by adapting the kernel arguments which are located in the `profile.yaml` file associated with our diskless OS image. As an example, the following is the kernel arguments for the compute servers that was used in the Lenovo Benchmarking cluster:

```
# cat /var/lib/confluent/public/os/rocky-9.6-cn5k-diskless-v1/profile.yaml
label: Rocky Linux 9.6 (Blue Onyx) (Diskless Boot)
kernelargs: clocksource=tsc tsc=reliable confluent_imagemethod=untethered \
hfi1.use_bulksvc=Y hfi1.num_sdma=8 hfi1.num_vls=4 hfi1.num_user_contexts=0,208 \
 initcall_blacklist=algif_aead_init
```

- To use BTS, we set `hfi1.use_bulksvc` to `Y`

- To ensure there are enough contexts available for BTS to start, we limit the number of user contexts to 208 using `hfi1.num_user_contexts=0,208` (it is 0,208 because we are using “port” 2)

- Additionally `num_sdma=8 num_vls=4` are set as optimal options for HFI driver as according to the CN5000 Performance Tuning Guide [2].

Then from within the booted OS, the user can choose to use BTS at runtime with the environment variable,

```
FI_OPX_HFISVC=1
```

Depending on the specific application and workload in question, this may end up improving or hindering the measured performance, either beyond a single node or past a certain point at scale.

3.2.3 SDMA

System direct memory access (SDMA) is another option that can be toggled to improve the performance of certain applications and workloads. With `num_sdma=8` set in our kernel options, SDMA will be used. The user can choose to disable SDMA at run time by setting:

```
FI_OPX_SDMA_DISABLE=0
```

4 Application Performance Review

In this section, we examine the scaling characteristics of HPC workloads whose behaviour is well understood in the HPC community. We begin with workloads that are known to scale efficiently (nearly linear) as the number of nodes increases, followed by workloads that exhibit sub-linear scaling or reach a performance plateau as the number of nodes increases. For each workload, we summarize the runtime configuration and tuning parameters that delivered the best performance.

4.1 Approximately Linearly Scaling Workloads

4.1.1 HPL (High Performance Linpack)

Although HPL is a synthetic benchmark rather than a real-world HPC application, it serves a valuable baseline for assessing scaling efficiency across the CN5000 Omni-Path fabric.

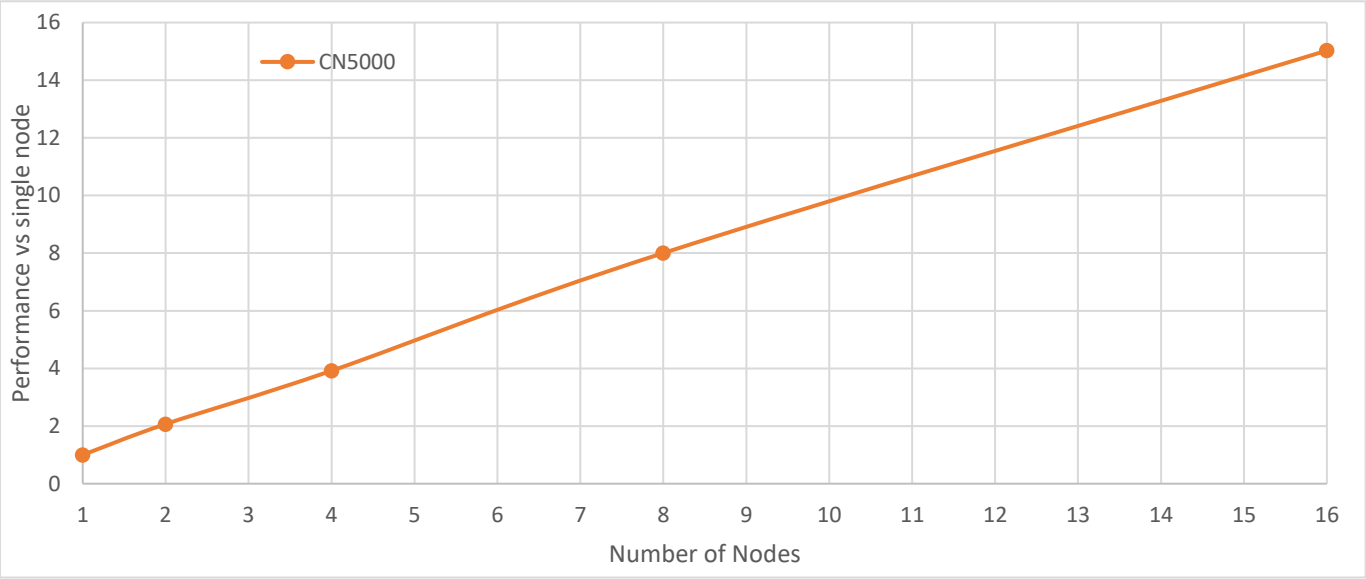

Figure 2: HPL scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

Figure 2 presents the relative performance of HPL compared to the measured single node performance versus number of nodes to showcase the linear scaling up to 16 nodes. For this benchmark we ran the Intel optimized release of HPL provided with the MKL libraries as part of OneAPI. The benchmark was executed with one MPI rank per socket so that there were two MPI ranks used per node. With only a couple of MPI ranks used per node, we didn't expect SDMA or BTS to make a difference in the results presented and when testing we confirmed this when seeing a sub 1% difference between the results with the different options, well within the statistical variation of the test.

4.1.2 CAE Workload Showcase - OpenFoam Car 300 million

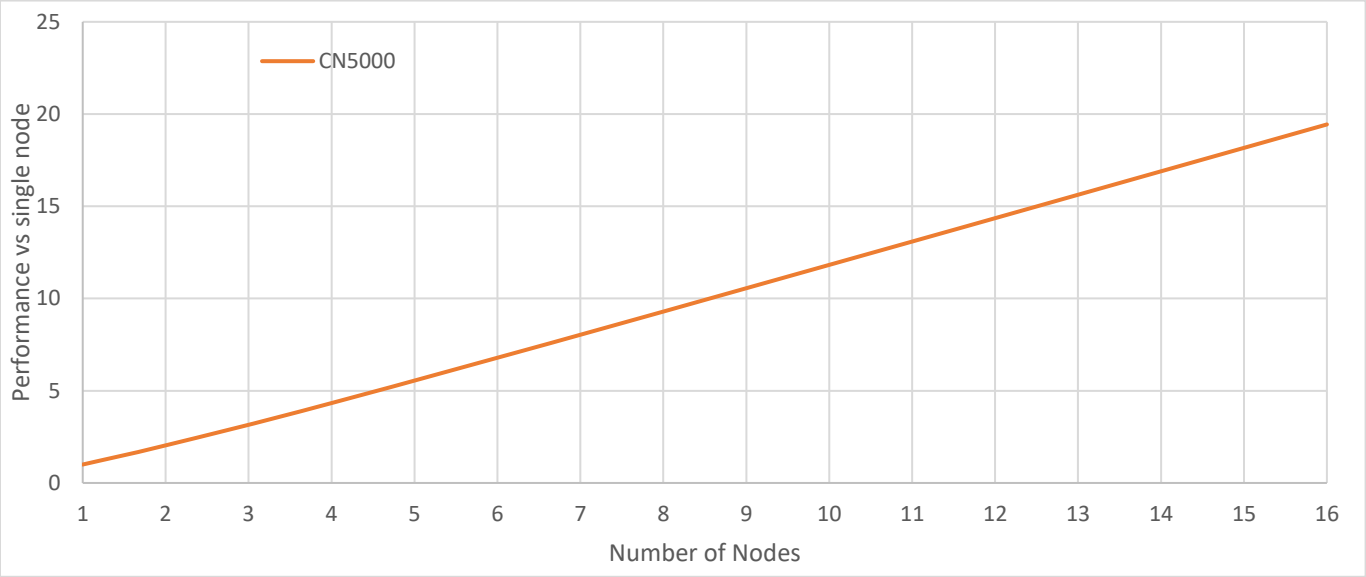

Figure 3 OpenFoam Car 300 million test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

OpenFoam is a free and open source CAE/CFD software maintained by OpenCFD which is available at https://www.openfoam.com/. Figure 3 presented results with a production scale OpenFoam automotive workload with ~300m cells which was chosen in order to assess application scalability up to 16 nodes when running in MPI mode with 256 tasks per node. For this workload we found that the best results were obtained with both SDMA and BTS disabled, where the favourable decomposition on the multimode setup allowed the code to scale super linearly.

4.1.3 CFD Workload Showcase - Siemens STAR-CCM+ LeMans Car 100 Million

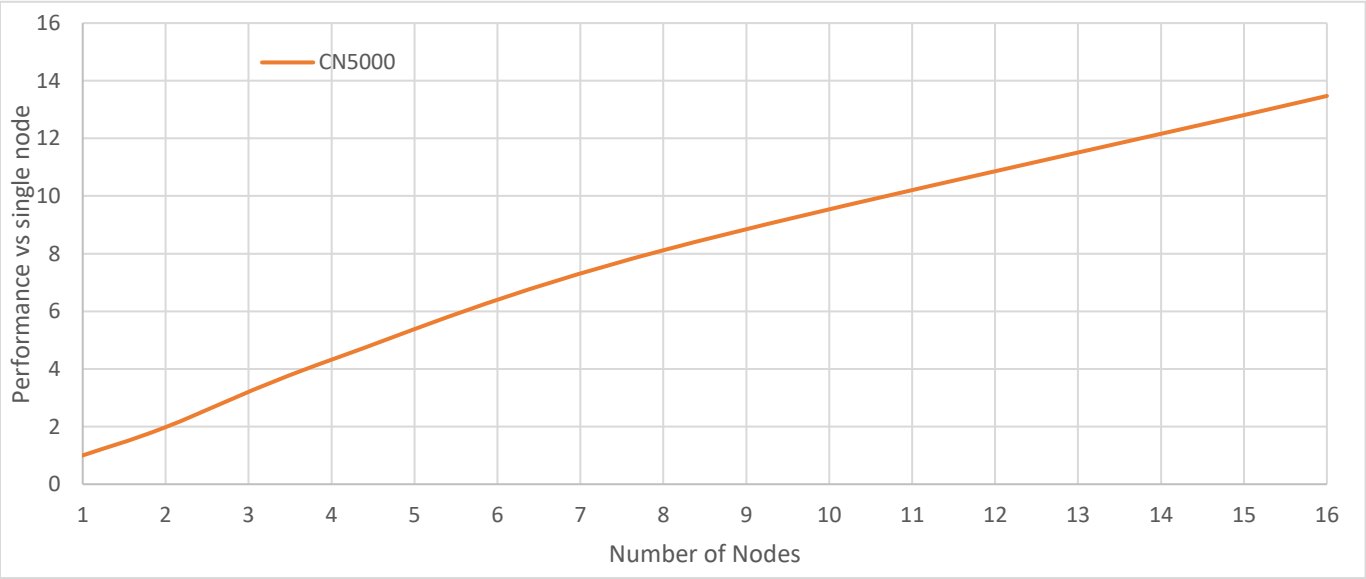

Figure 4 Siemens STAR-CCM+ LeMans Car 100M test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

STAR-CCM+ is a commercial CFD code provided by Siemens and available at https://www.siemens.com/en-gb/products/simcenter/fluids-thermal-simulation/star-ccm/. Figure 4 presents the LeMans Car 100m test case as an

example of a larger CFD workload running on up to 16 nodes. With this workload, we observed minimal performance differences between SDMA and BTS configurations when running on 1 to 4 nodes. However, at larger scales of 8 and 16 nodes, the best performance was achieved with BTS and SDMA enabled.

4.1.4 Molecular Dynamics Workload Showcase – NAMD2.15 Memory Optimised 20STMV

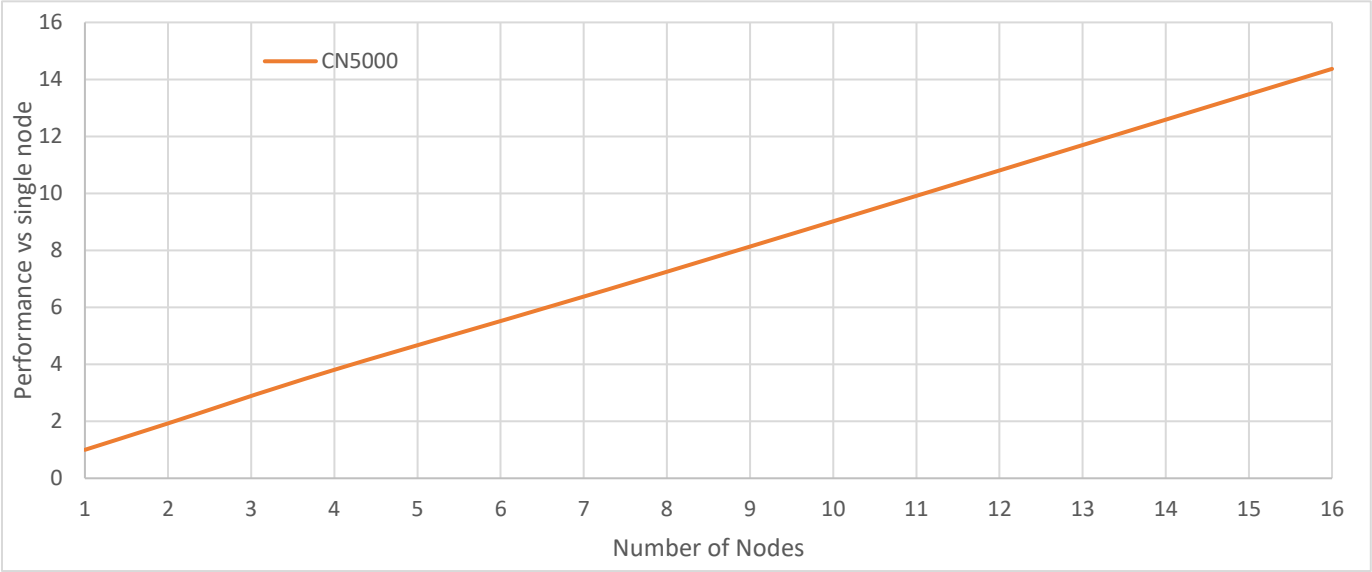

Figure 5 NAMD 20STMV test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

NAMD (Nanoscale Molecular Dynamics) is a code used for molecular dynamics simulation maintained by the US National Institutes of Health as part of the University of Illinois at Urbana-Champaign. Is it available from https://www.ks.uiuc.edu/Research/namd/. For this workload we used the memory optimised build of NAMD, which included preparing the dataset according to the README in https://www.ks.uiuc.edu/Research/namd/utilities/stmv_sc14.tar.gz. Figure 5 presents the scaling performance of NAMD with the 20STMV benchmark which simulates over 20 million atoms of the Satellite Tobacco Mosaic Virus. This workload scales nearly linear up to 16 nodes and having both SDMA and BTS disabled gave the best performance for all test points between 1 and 16 nodes.

4.1.5 Weather Workload Showcase WRF Maria 1Km

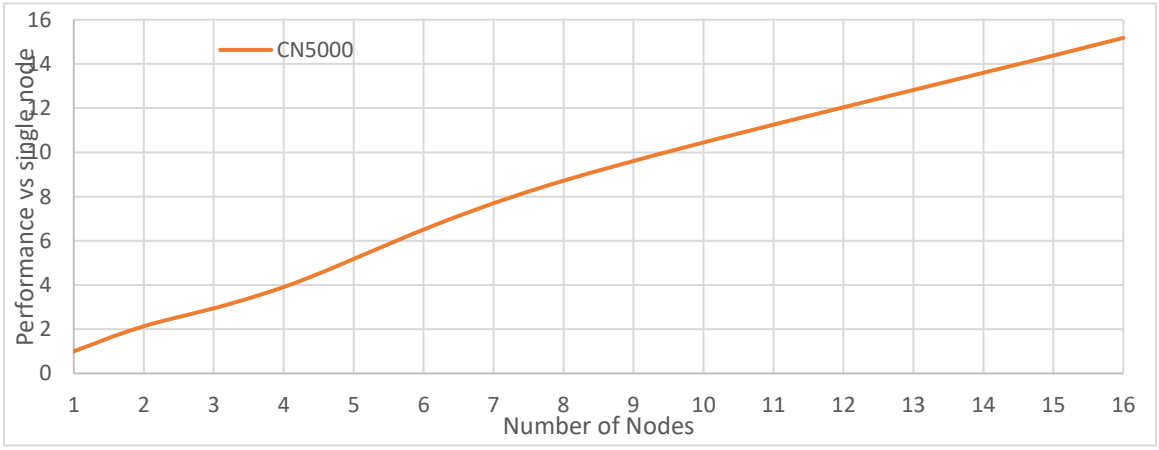

Figure 6 WRF Maria 1km test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

The Weather Research and Forecasting code (WRF) is a free numerical weather prediction code provided by NCAR the U.S. National Center for Atmospheric Research. The code is available at https://www.mmm.ucar.edu/models/wrf. Figure 6 presents the scaling performance of the Hurricane Maria 1km benchmark testcase[5]. We observe linear scaling up to 16 nodes. The results presented were measured with both SDMA and BTS enabled.

4.2 Sub-Linearly Scaling Workloads

In this section we examine workloads that are known to scale sub-linearly as the number of node increase, often reaching a plateau in performance beyond a certain scale. As in the previous section we use real-world workloads spanning the four major areas of HPC to illustrate these scaling characteristics.

4.2.1 CAE Workload Showcase - OpenRadioss Taurus

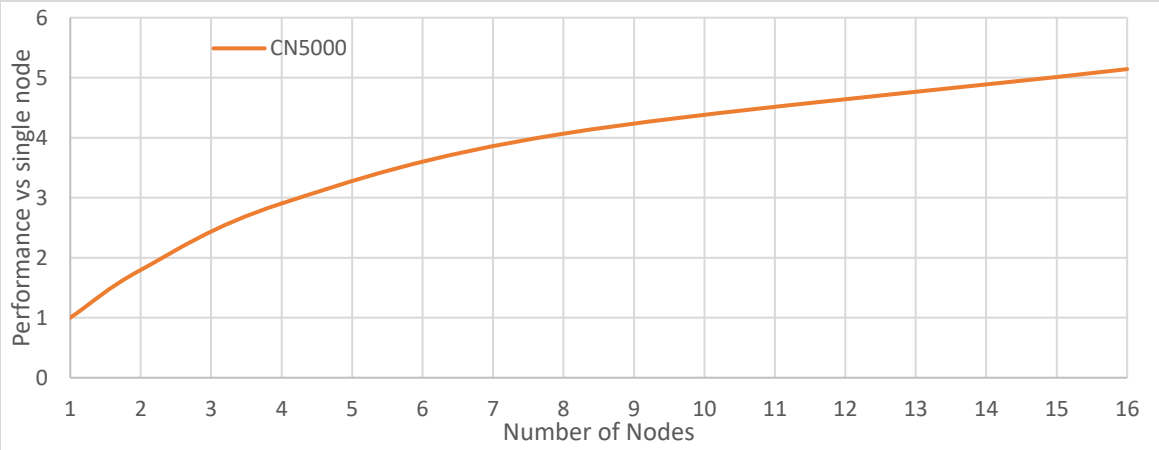

Figure 7 OpenRadioss Taurus 10m test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

OpenRadioss is a free crash simulation software maintained by Altair Engineering. It is available at https://openradioss.org/. The Taurus benchmark[6] consists of a 10 million element model of a Ford Taurus which we use in Figure 7 to study the scaling of OpenRadioss on Thinksystem SC750 V4 with Intel 6980p Processors. For this workload, the scaling performance reduces to sublinear at around 3 nodes and continues to degrade as the number of node increase. The best performance was achieved using SDMA enabled and BTS Disabled.

[5]available from NCAR website https://www2.mmm.ucar.edu/wrf/users/benchmark/v44/benchdata_v44.html

[6] Available from: https://openradioss.org/models/

4.2.2 Chemistry Workload Showcase - Quantum Espresso GRIR

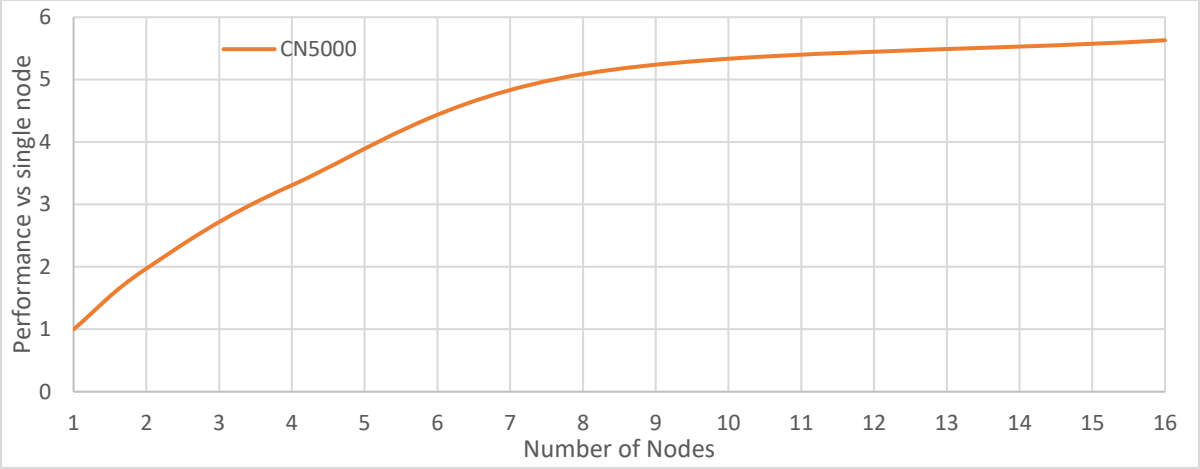

Figure 8 Quantum Espresso GRIR test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

Quantum Espresso is a quantum chemistry / electronic structure theory code to model materials at the quantum level. It is maintained by the Quantum Espresso Foundation and available at https://www.quantum-espresso.org/. The GRIR benchmark[7] simulates a carbon-iridium complex $C_{200}Ir_{243}$ and we use this benchmark in Figure 8 to study QE performance at scale. Linear scaling breaks around 4 nodes for this workload and continues to worsen to 16 nodes resulting in just under 6x performance uplift from the single node result. For this workload, we found having SDMA on and BTS off gave the best performance, except for 16 nodes where having BTS enabled was the best option.

4.2.3 CFD Workload Showcase - Siemens STAR-CCM+ VTM Uhood 64M

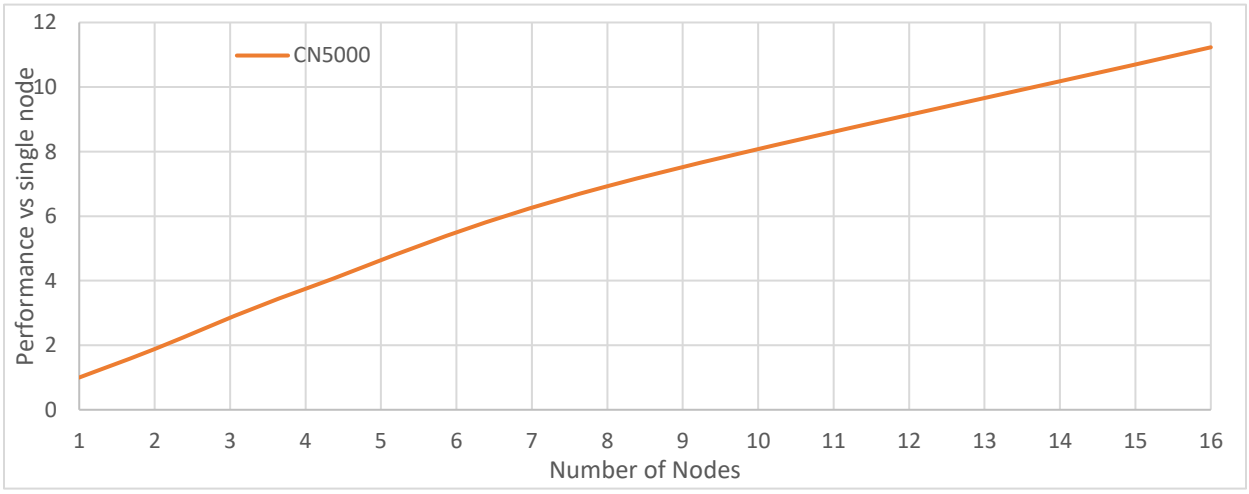

Figure 9 Siemens STAR-CCM+ VTM Uhood 64M test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

The VTM (Vehicle Thermal Management) Uhood (Under-hood) model with 64 million cells was used here as an example of

[7] Available from https://github.com/QEF/benchmarks/tree/master

a CFD workload which scales sub-linearly and the scaling behaviour is presented in Figure 9. Beyond around 4 nodes, the scaling of the workload begins to drop off until at 16 nodes the performance is roughly 11x that of a single node. Similar to the other STAR-CCM+ workload in the previous section, we see that for 1 to 4 nodes the performance was independent of BTS and SDMA options but for 8 and 16 nodes the best performance was observed with both SDMA and BTS enabled.

4.2.4 Molecular Dynamics Workload Showcase – NAMD3 STMV

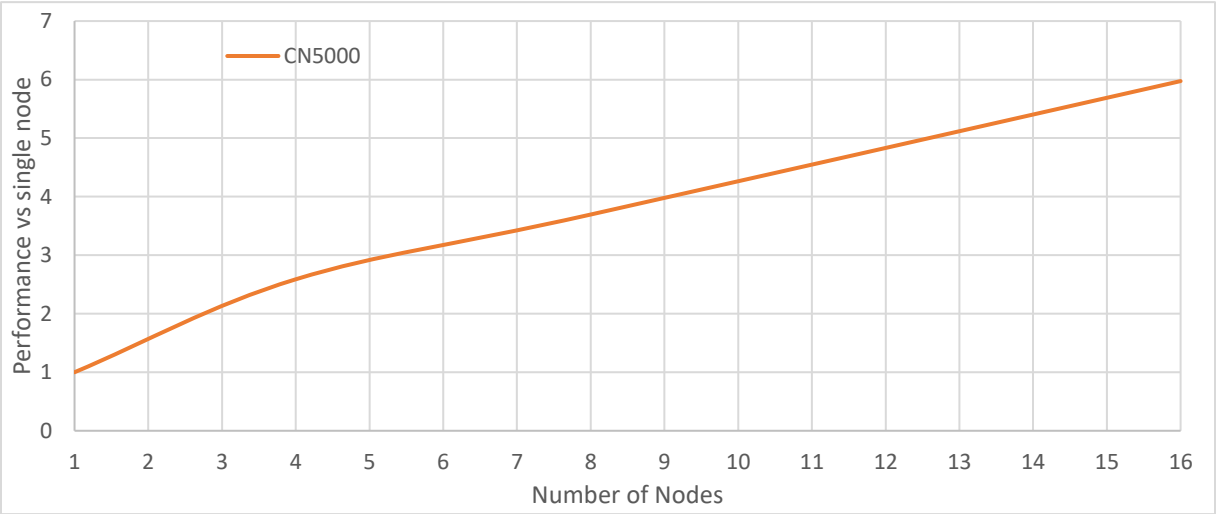

Figure 10 NAMD STMV test case scaling scaling on up to 16 nodes of ThinkSystem SC750 V4 with Intel 6980P processors and CN5000 Omni-Path.

NAMD (Nanoscale Molecular Dynamics) is a code used for molecular dynamics simulation maintained by the US National Institutes of Health as part of the University of Illinois at Urbana-Champaign. Is it available from https://www.ks.uiuc.edu/Research/namd/. In this case we used the regular non-memory optimised version of the NAMD code. Figure 10 presents the scaling performance of NAMD with the STMV benchmark which simulates over a million atoms of the Satellite Tobacco Mosaic Virus. For this workload, the scaling behaviour is sub-linear and only reaches 6x the single performance. We contrast this behaviour with that of the previous NAMD workload examined in Figure 5, which was able to scale nearly linear up to 16 nodes. This is because of the improved communication/computation ratio when using a much larger test case. Similar to that previous NAMD workload, having both SDMA and BTS disabled gave the best performance all test points between 1 and 16 nodes.

4.3 Summary of Application Best-Recipes

In Table 2 and Table 3 summarizes the testing environment, beginning with the system configuration and followed by the application and runtime settings used for each workload.

Option	Setting used
OS	Rocky Linux 9.6
Linux Kernel	5.14.0-570.58.1.el9_6.x86_64
UEFI – Workload Profile	“HPC Mode” with the additional changes below
UEFI – Additional Options	UEFI.Processors_UncoreFrequencyScaling "Enabled" UEFI.Processors_VirtualNuma "Disabled" UEFI.Processors_SNC "Enabled"

	UEFI.Processors_L0p "Enabled"
OPXS Version	12.1.1.0.18

Table 2: System Settings needed to replicate the performance showcased in this review.

Application	Version	BTS	SDMA
HPL	2023.0.0	N/A	N/A
OpenFoam – Car 300m	2512	Disabled	Disabled
StarCCM – LeMans Car	20.02.007-R8	Enabled	Enabled
OpenRadioss – Taurus	20260319	Disabled	Disabled
QE – GRIR	7.5	Enabled	Enabled
WRF – Maria 1Km	4.7.1	Enabled	Enabled
StarCCM+ - VTM Uhood	20.02.007-R8	Enabled	Enabled
NAMD Memory Optimised	3.0.2	Disabled	Disabled
NAMD STMV	3.0.2	Disabled	Disabled

Table 3 Toolchain and runtime options which gave the best performance at 16 nodes for each of the workloads presented in this section.

4.4 Further Studies and Related Work

In addition to the application results present here, there are two additional releases on LenovoPress diving into specific areas of interest:

- Lenovo CAE Reference Architectures Powered by Cornelis CN5000 Omni-Path [3]: Outlines a Reference Architecture for HPC tailored specifically to Computer Aided Engineering Workloads
- Installation and Best Practices for Implementing Cornelis CN5000 Omni-Path on Lenovo ThinkSystem Servers [1]: Outlines the best practices for setting up Omni-Path on ThinkSystem and includes some discussions on microbenchmarks recommended for bring up.

5 Conclusions

In this paper, we presented a validated Reference Architecture for scalable HPC workloads using Lenovo ThinkSystem servers and Cornelis CN5000 Omni-Path networking. Through a diverse set of representative HPC applications spanning engineering, scientific computing, molecular dynamics, weather forecasting, and computational chemistry, we demonstrated how the architecture delivers efficient scaling from a single node to a 16-node deployment. The results highlight both workloads that scale nearly linearly and those whose performance becomes increasingly communication-bound at larger node counts, providing practical insight into the behavior of real-world HPC applications on the CN5000 Omni-Path fabric. The scaling behaviour of the tested workloads met expectation which allows customers to get the most out of Lenovo ThinkSystem servers when using the Cornelis CN5000 Omni-Path network.

Beyond presenting application performance results, this study also examined the impact of key network and runtime tuning parameters on scalability. The recommendations provided throughout this paper can serve as a starting point for optimizing workloads not explicitly covered by this study. While the optimal configuration remains application-dependent, we observed that features such as Bulk Transfer Service (BTS) often provide performance benefits at larger scale and should be considered when tuning communication-intensive workloads. Future CN5000 Omni-Path software releases are expected to introduce additional enhancements, with updated guidance continuing to be documented through Lenovo EveryScale Best Recipe publications.

The reference architecture is designed to scale beyond the validated 16-node configuration. By leveraging the available ports on a single CN5000 Edge Switch, up to three N1380 Neptune enclosures can be connected within a single-switch topology, enabling a full-rack HPC deployment. Larger environments can be accommodated through multi-tier fabric architectures utilizing additional CN5000 switching infrastructure, allowing the design to grow from departmental clusters to larger-scale HPC installations while maintaining a common architectural foundation.

References

[1] C. Elrick, “Installation and Best Practices for Implementing Cornelis CN5000 Omni-Path on Lenovo ThinkSystem Servers,” July 2026. [Online]. Available: https://lenovopress.lenovo.com/lp2474-installation-and-best-practices-cornelis-cn5000-omni-path.

[2] Cornelis Networks, “CN5000 Performance Tuning Guide V3.0,” 2026. [Online]. Available: https://customercenter.cornelis.com/.

[3] K. Dean and K. Kutzer, “Lenovo CAE Reference Architectures Powered by Cornelis CN5000 Omni-Path,” 2026. [Online]. Available: https://lenovopress.lenovo.com/lp2372-lenovo-cae-reference-architectures-powered-by-cornelis-cn5000-omni-path.

Authors

Conor Elrick is a HPC/AI Performance Engineer in the ISG Offerings Group at Lenovo working as part of the HPC/AI Solutions Architect Team. Conor does setup and performance benchmarking on compute servers, networking infrastructure and storage solutions to push hardware to its limits and get the best results for real world scientific workloads. He has been at Lenovo for over 3 years. Conor earned his PhD in Theoretical Particle Physics from the University of Edinburgh in 2024 and a Master's Degree and Bachelor's Degree in the same field before that, with a focus on physics simulations on compute systems.

Kevin Dean is the Senior Manager of the HPC/AI Solution Architect Team in the ISG Offerings Group at Lenovo. Kevin oversees the strategy and processes for HPC and AI solutions in this position, contributes his knowledge of HPC/AI application performance, and serves as the CAE/manufacturing architect. Kevin has over 20 years' experience in HPC, AI, and computational engineering, including nine years at Lenovo focused on HPC/AI performance and solution architecture, and 12 years in aerodynamic design and CFD for the U.S. defense and automotive racing sectors. He earned his Master's Degree in Aerospace Engineering at the University of Florida, following a Bachelor's in the same field from Virginia Polytechnic Institute and State University.

We would like to thank:

- Cornelis Networks team, including Paul Stasurak, John Swinburne, Balaji Srinivasan, James Erwin, Pavel Dobrinskiy, and Tyler Cruise for their valuable guidance, technical input, and validation of the installation and setup.

Trademarks and special notices

© Copyright Lenovo 2026.

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®
Neptune®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

Intel® is a trademark of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

IBM® is a trademark of IBM in the United States, other countries, or both.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.