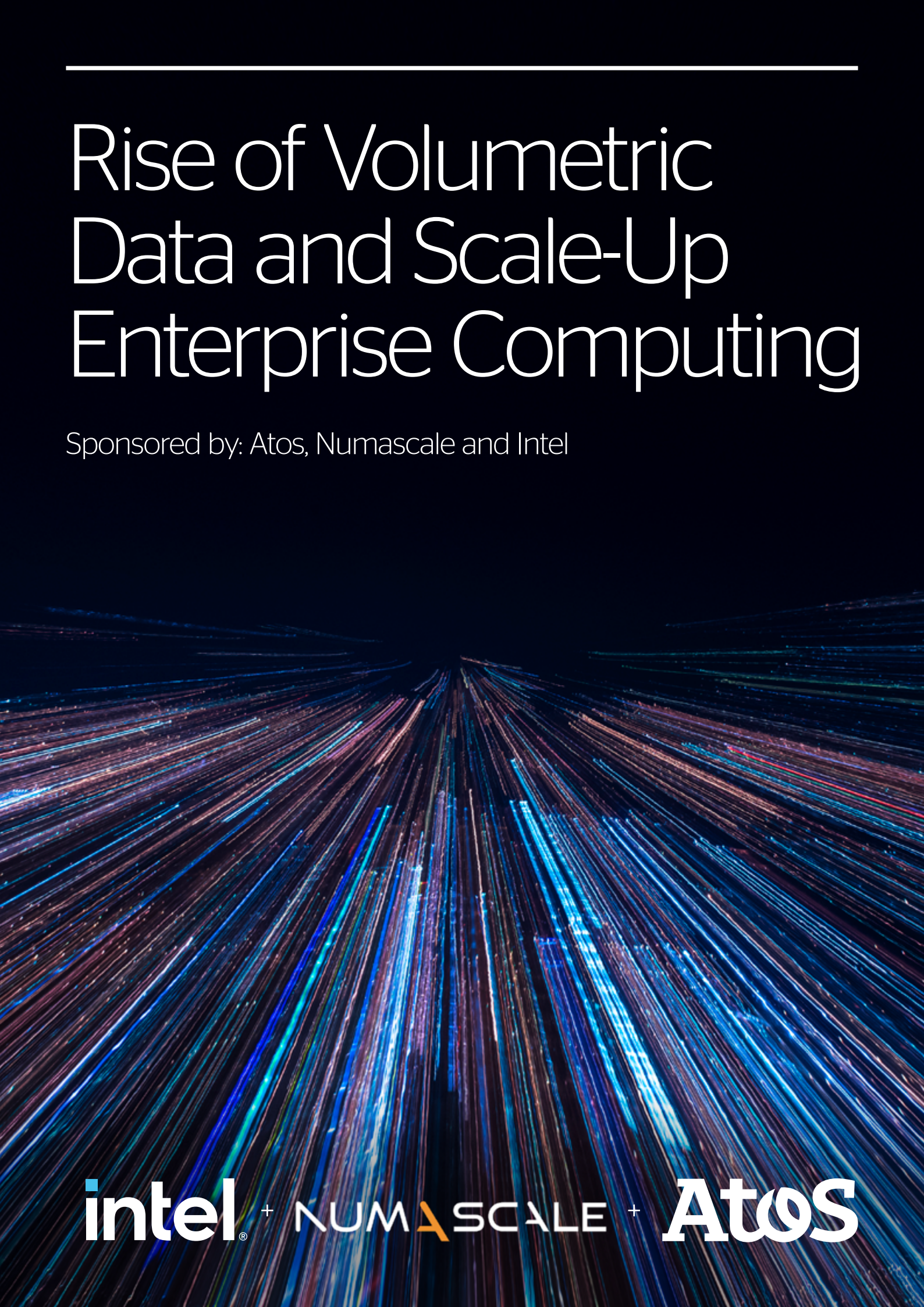
Rise of Volumetric Data and Scale-Up Enterprise Computing

Sponsored by: Atos, Numascale and Intel



Mise en Scene

This paper captures the continued collaboration among Intel, Atos and Numascale to enable cost effective Scale-Up ecosystem on x86 server platform.

Intel is the technology provider (QuickPath Interconnect, Ultra Path Interconnect), Atos is the server platform provider (BullSequana S), Numascale is the node controller architecture provider (xNC).

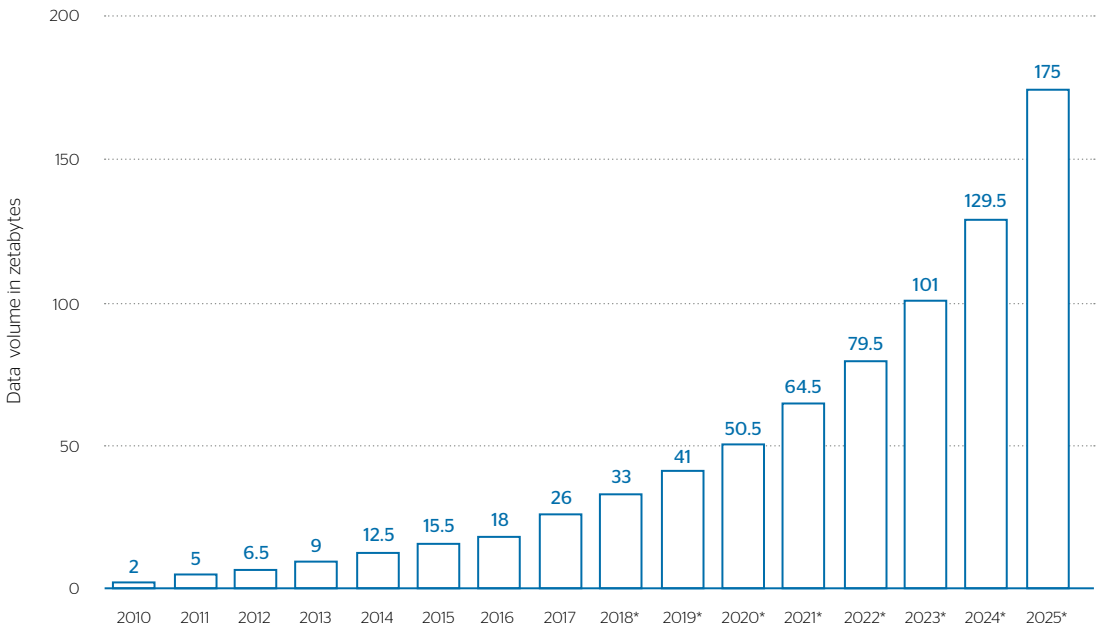
The volumetric data is rising fast, how to process 50+ Zettabytes of today's data through efficient computing is not a small task to say the least.

This paper describes the plan, methodology and the roadmap to address this critical computing need. It will introduce Atos platform with external Node Controller (xNC) architecture, jointly developed with Numascale and creating customer value from one generation of Scale-Up server system to the next.

The Rise of Volumetric Data Processing: AI, Machine Learning and Natural Language

Depth.Scale.Latency.Gravity.Volume

As the planet hurls through the galaxy, satellites circle the globe, we humans continue to consume, build, model and render data at unprecedented levels.



The current data forecast for 2025 is almost 175ZB (Zettabytes). The forecasts have been traditionally very accurate within 2-5% over long-time horizons in my experience. Our species requires this data for so many functions from finance, healthcare, education, research, communication, transportation and entertainment to name a few. We share our experiences, our health, wealth and well-being. Our lives, biometry, births, deaths and legacy. We have become our own videographers through very capable globally sourced smartphones nearly 17 years after their first introduction. By 2025, nearly 3.7B people will use this technology to access the internet on a regular basis, according the CNBC. We have developed the capability to communicate across the globe in near real time. In the last 20 years we have created more devices to consume, create, alter, and re-imagine data than any industry analyst or pundit could have possibly imagined.

One-dimensional (1D), two-dimensional (2D), three-dimensional (3D) and four-dimensional (4D) data creation have changed the compute architectures we have to build. Dynamic new math libraries and artificial intelligence (AI) instructions will provide our customers with capabilities we could only imagine a decade ago. A new era of design is required...

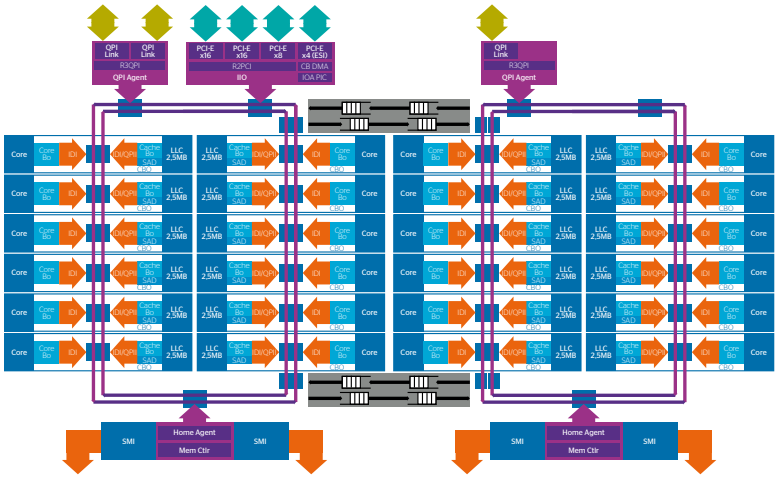
Why does this matter? Why should we care? 3D and 4D data has already begun to lead our fight for survival with Covid-19 and healthcare professionals. These models can then be shared, anonymized and rendered to allow doctors, nurses, medical technicians and healthcare companies to closely examine the results. There are many active teams across the globe working tirelessly to continue to find a cure for years. The depth of the data matters. The more data sets, more points of reference, more bytes, more bits all help when using AI data models and creative Data Scientists to build next generation models to understand virology, magnetic resonance and telehealth services. The depth of the data matters, when it is your life.

Non-Uniform Memory Access (https://en.wikipedia.org/wiki/Non-uniform_memory_access) was originally designed to provide single operating system instances to scale beyond single socket central processing units (CPU). This work pioneered by Atos/Bull, Sequent/IBM, DEC and Intel in the mid-1990's has become a foundation of scaling compute architectures today. From CPU to Rack Design, the principles of scaling and Non-Uniform Memory Architectures (NUMA) can be found. Scale is critical to provide greater results, larger databases and more compute, memory, interconnect and network resources.

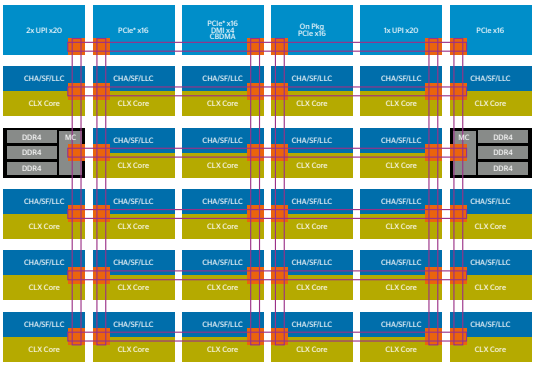
Intel Xeon System on a chip (SoC) architecture

Continued emphasis on modularity & balanced performance scalability

Intel Xeon Processor E7 (24 cores)



2nd Gen Intel Xeon Scalable Processor (28 cores)



SAD: Source Address Decoder
QPI: Intel QuickPath Interconnect
IIO: Integrated I/O
PCU: Power Controller Unit
Ubox: Processor Utility Box

CHA – Caching and Home Agent
SF – Snoop Filter
LLC – Last Level Cache
CLX Core - 2nd Gen Intel Xeon Scalable Processor core
UPI – Intel UltraPath Interconnect

Example of -2-socket Intel Xeon 6258R (28 cores, 2.7GHz) SNC-OFF	Memory and cache latencies w/Intel Memory Latency Checker (MLC)
L3 hit cache (same socket)	20.2 ns
L3 hit cache (remote 2nd socket)	180 ns
DRAM hit memory (same socket)	80 ns
DRAM hit (remote 2nd socket)	138 ns

A bumble bee flap has been recorded at approximately 5 milliseconds. Most of us today around the world experience our broadband internet speeds at between 30-100 milliseconds of latency. A bumble bee is more latency-aware than most in our species. Latency is critical, few places more critical than system architecture and latency more important than in the processing of volumetric, visual and AI data. Each part of the architecture must address latency from instructions, through memory, networking, interconnect and transport. Each leg of a "Bit's journey" must be as latency-free and optimized for performance to insure a balanced compute architecture. We have invested across Intel® Xeon® Scalable platform generations with instructions to reduce latency in virtualization, cache (Intel® Resource Director Technology), system check, security, high-bandwidth memory and management tools.
Each generation of Intel Xeon Scalable processors has been optimized to perform with our new memory architecture known as Intel® Optane™ Persistent Memory and the latest DDR technologies.
With the capabilities to scale from a single-socket Intel Xeon Scalable system up to 32-Sockets per physical server node with BullSequana S series platform customers all over the world will enjoy one of "the industry leaders" in scalability, NUMA and latency optimizations the world has ever seen.

Theoretically, NUMA system and micro-architectures can scale almost infinitely with linear graph performance characteristics, in the lab. Time, research, failure, success, fact and re-investment have all proven this theory to be false. Gravity often brings many theoretical mathematical and scientific discoveries back to earth. Gravity also provides us with insights how to defy and manipulate its principles to survive in space.
As data growth, capture and contextually aware applications begin to drive the transformation of industries...data is more than bytes and bits or bits and bytes. Data gathers value over time, when partnered with the right algorithms, databases, governance engines and toolsets. According to DOMO, 2.5 quintillion bytes a day of data is created by humanity and the devices we control, daily. With global adoption of devices growing at 7.5% a year from a 3.7B users base, the "gravity", scope and scale of our "data opportunity" becomes clear. Architectures must be designed for all types of usage models, atmospheric conditions, across a wide range of industries, enabling the broadest ecosystem of applications insuring consistent performance over the platform life cycle. 28-Core 2-32-Socket solutions with Intel Xeon Scalable processors, Intel Optane Persistent Memory, and Intel SSD technology with current Atos BullSequana S platform are at the core of our vision to build a more secured, scalable, high bandwidth, NUMA-optimized platforms for the next decade.

If past is prologue and the future resembles the last decade of data volume, volumetric data, user generated images and new data usage models for AI then the data forecasts beyond 2025 will be eclipsed. Beyond 2025 we will be living on a planet that requires dynamic platforms and technologies to manage 100's and 1000's of Zettabytes annually, doubling each year from 2025-2030.

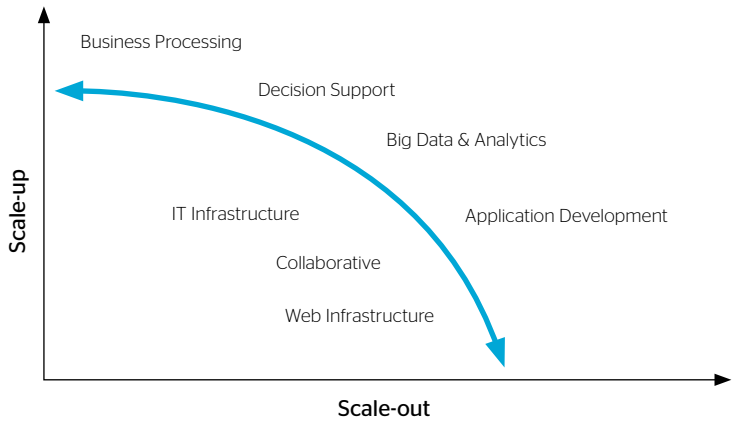
Introduction of Atos Platform and joint Numascale xNC Architecture

Scale-Up or Vertical Scaling (Stacking up servers vertically on a single compute node)

Scale-Up is done by adding more resources to an existing system or to a compute node to increase the performance of the node. Scaling up provides a large memory footprint across multiple CPU sockets (8S, 16S, 32S, 64S, etc.). A higher number of CPU sockets will then constitute a larger system with correspondingly larger memory to handle applications with large memory and compute requirements. Scale-Up uses extensively NUMA architecture, as this large memory, distributed close to the CPU, is accessed with many different latencies. Among the applications that benefit from Scale-Up systems are transactional databases and advanced analytics, like SAP HANA, scientific simulations, Real-time streaming, Artificial Intelligence (AI), etc. A common requirement for all of those applications is to have access to large amounts of memory across the CPUs (sockets) in a cache coherent manner. Conceptually, a large number of CPUs are vertically connected with node controllers to achieve 2X performance and more.

Scale-Out or Horizontal Scaling (Distributed connection of servers over many compute nodes)

Scale-Out is a distributed computer architecture. Typically, a compute node consists of 2-4 CPU sockets.
A number of such nodes are then horizontally connected in non-cache coherent way through a network fabric.
The memories are not visible across nodes. It is not possible to obtain a single memory image (across sockets) as we get in Scale-Up systems. This will limit throughput and capacity for all applications that are memory "hungry" and compute intensive.



Evolution of Scale-Up Technology at Bull and Atos:

Today, Bull, now part of Atos, is one of the very few developing Scale-Up and NUMA servers, based on Intel processors.

Bull teams started in the 1980s to build its long experience and expertise on memory and cache coherency protocols by designing and developing proprietary multi-processor systems (DPS7 and DPS8). Those systems first connected several processors to memory controller through a snoop-bus, then to a shared level of cache in front of the memory controller, providing at this time a uniform access to the memory (UMA).

Then, Bull teams developed in the 2000s their first NUMA server. That one was based on Intel Itanium processors. For this purpose, Bull took benefit of this expertise to define a directory-based extension of the Intel coherency protocol and to implement it in a dedicated ASIC (called B_SPS). This ASIC is an eXternal Node Controller (xNC), and has been used at the heart of the new 16-processor server. The basic Intel platform connected four processors and two IO controllers to one Memory Controller to build a 4-processor system. Two instances of the B_SPS node controller allowed to interconnect two Intel IO controllers and four Intel Memory controllers and so allowed to build a 16-processor server with the requested bandwidth. Each B_SPS provides also two external ports to build a 32-processor system when connecting directly two B_SPS together and an evolution has been studied to build a 64-processor system by connecting four B_SPS in a ring topology. With this server, launched on the market under "NovaScale" branding, Bull entered the High-Performance Computing area, and installed in 2005 at French CEA the supercomputer TERA 10, which overpassed 10 Teraflops.

Later, Bull developed new generations of external node controllers and servers, for any new generation of Intel Xeon scalable platform. Such servers targeted mainly enterprise market with high memory capacity, such as database in memory, and for this purpose provided maximum memory capacity, many PCIe slots and high levels of reliability and serviceability.

For High Performance Computing, they addressed only pre-processing and post-processing phases, as Bull introduced blade servers to take advantage of the applications parallelism.

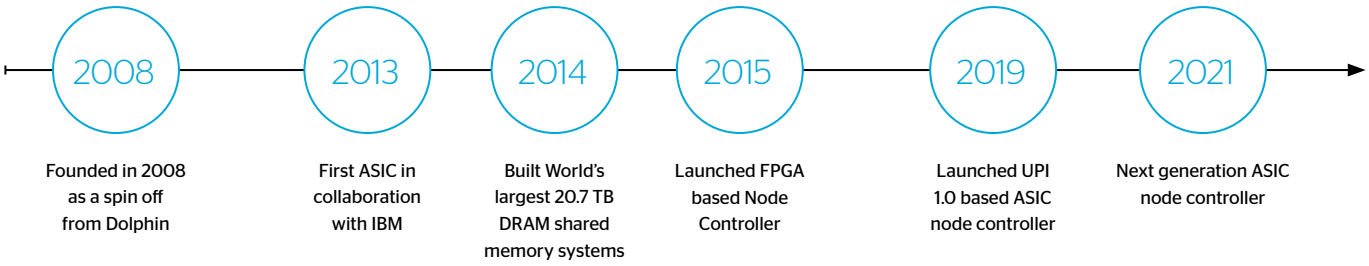
The architecture of the external node controller ASIC had to adapt to the new Intel protocols and to the evolution of the Intel reference platform. For example, in previous platforms, the memory controller was embedded into the processor socket, which provided four QPI 1.0 point-to-point links, three of them to interconnect four processors in an all-to-all topology, and the fourth one to connect to the IO controller. In "Bullion" named servers in the 2010s, this fourth QPI links was connected to a new ASIC (BCS) connected itself to the IO controllers. Bull built a 4-processor drawer where one BCS interconnected four processors and two IO controllers. BCS provided also six proprietary extended coherent ports implemented with cables, in order to build symmetric multiprocessors with two drawers (8 processors), three drawers (12 processors) and four drawers (16 processors), providing up to 2TB of mail memory. Then, starting in 2014, the Intel architecture came again different with previous Intel Xeon processors. On this platform, the IO controller was embedded inside the socket and the number of QPI 1.1 links limited to three. Bull took advantage of this evolution to design a smaller ASIC (BCS2) and to improve the scalability of the server on a 2-socket basis. This new BCS2 provided two QPI 1.1 connections and seven proprietary extended coherent serial ports. It was connected to two sockets on a motherboard hosted in a 2-socket drawer. Such design provided the opportunity to build the scalable symmetric multiprocessor "Bullion S" from two drawers (4 sockets) to eight drawers (16 sockets) with all-to-all cable connections between up to 8 BCS2. Bullion S provides up to 24 TB of main memory.

The Intel architecture did not change with 1st and 2nd Gen Intel Xeon Scalable platforms in 2017, but the coherency protocol evolved deeply with the introduction of UPI for connection between sockets. With this new platform Atos developed the "BullSequana S" server and improved its design by providing a same 2-socket drawer basis for glueless configurations up to eight sockets (the maximum supported by Intel UPI protocol) and from ten to thirty two sockets for configurations with node controllers. The main interest was to provide entry level 4-socket and 8-socket configurations.

In order to achieve this flexibility, the interface of the 2-socket drawer could no longer be an Atos proprietary interface but should be a UPI interface. It was the first implementation of UPI on cable, to connect either up to eight sockets together in a glueless topology or to connect sockets to an external node controller drawer containing the new generation of ASIC (xNC). The concept of one ASIC connected to two sockets has been kept, but one socket could be connected to two instances of ASIC in dual rail configuration, in order to double the bandwidth between sockets in remote drawers.

Evolution of Scale-Up Technology at Numascale:

Numascale® Company was founded in Oslo in 2008 as a spin-out from Dolphin Interconnect Solutions. Numascale technology has roots back to Norsk Data's and Dolphin's Scalable Coherent Interface (SCI) products. Numascale vision is to "Enable cost-effective Scale-Up server ecosystem and the mission is to "Delight our Partners and Alliances delivering best in class cache coherent node controller technology in Scale-up computing." Numascale engineers developed technologies in several commercial products in the past, such as: Convex Exemplar (now HP Superdome), Data General Aviion (Acquired by EMC), SunCluster (Sun Microsystems). Numascale core xNC architecture is cache-coherent IO technology agnostic. Numascale developed products on Intel QuickPath Interconnect and multiple generations of Intel Ultra Path Interconnect technologies and other x86 architectures.



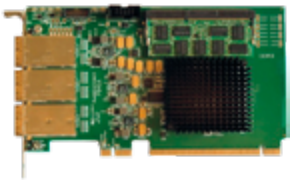
Past Scale-Up Numascale Products:

HT ASIC NC

HT FPGA NC

Cable

HTX Card



5000 Cores, Single Image OS

- 108 nodes
- 6x6x3 Torus
- 5184 CPU cores
- 58 Teraflops
- 10TB/s Memory BW

Common Atos-Numascale Platform development:

Numascale and Atos jointly developed the cache coherent node controller (xNC) to build 8S, 16S, 32S scale-up servers. xNC provided two UPI 1.0 interfaces and eight external interfaces of Numascale IP. The bandwidth between sockets inside a 16-socket was optimal with an all-to-all topology. These external connections between xNC were routed inside the xNC drawer from 8-socket to 16-socket configurations, and between two drawers from 18S to 32S. BullSequana S in 16-socket configuration provides up to 48TB of main memory.

This Intel UPI 1.0 based external Node Controller (xNC) has been deployed with BullSequana S series Scale-Up servers. Future product with next generation UPI is targeted for 2021/2022.

Picking the right models and framework to scale

Although virtualization and cloud adoption have favored scale-out deployments, they are not well suited to business processing, big data and analytics (e.g. SAP HANA) workloads, requiring maximum resources to process vast amounts of data. These applications can take advantage of the large number of processors that are close together and the large memory capacities of the BullSequana S scale-up systems, allowing huge databases to be completely stored in-memory. These computational resources located near the data as well as the use of interconnects strongly contribute to provide better performance by eliminating a complex network mesh to connect nodes and latencies of memory access.

Furthermore, scaling up allows to reduce not only operational costs because of simplified management, reduced power and cooling costs, but also results often in TCO savings from lower licensing costs. For example, running Oracle database on a scale-out architecture using Oracle RAC clustering can be a much more expensive solution than with a scale-up configuration for which Oracle RAC is not required.

BullSequana S enterprise servers have been designed to host mission-critical memory-intensive applications with high throughput requirements, such as SAP HANA, Oracle Database or artificial intelligence (AI) applications. These data-driven applications with increasing amount of data can take advantage of the near-linear scalability, the performance and the reliability of BullSequana S servers that break through the limits set by 4-socket servers, frequently deployed in the datacenters.

BullSequana S uses a modular scale-up x86-based architecture allowing to start at 2 CPUs/64GB DRAM and scale-up seamlessly up to 32 CPUs/48TB of shared memory in 2-socket increments as a single system, which is unique on the market, thus meeting the most data-intensive and demanding workloads.

The module is the building block of BullSequana S servers. It contains a Compute unit with CPU, memory and some internal storage plus an optional Storage unit or GPU unit to customize and extend the system to match application requirements.

Each Compute unit takes 2U in a rack and includes:

- Two 2nd Generation Intel Xeon Scalable processors, with a large choice of models in terms of frequency, number of cores or power consumption, for the best of your applications
- Up to 24 memory DIMMs, i.e. a total of up to 3TB per compute module when using 128GB DIMMs
- Non-Volatile RAM (NVRAM) capabilities with the support of Intel Optane Persistent Memory (DCPMM) providing near-DRAM performance at a lower-cost. BullSequana S can run either with all DRAM or with a combination of DRAM and NVRAM. Furthermore, Intel Optane Persistent Memory can reduce considerably downtime with a much quicker reload of the data when the system restarts.
- Up to 8 disks and hot-plug PCIe blades.

GPU unit, for artificial intelligence and machine learning
This option allows to introduce up to 32 GPUs in a single server in a very flexible way, 2 GPUs per module. Real-time algorithms and machine learning will use this huge processing power to run.

Storage unit, for data-extensive needs
This optional unit can hold up to 12 SAS/SSD 2"5 disks; 4 NL-SAS high capacity 3"5 disks; 4 NVMe for high I/O throughput.
Thanks to this additional Storage unit, the capacity of each 2-socket module goes up to 20 disks in a 2U form-factor, and more than 2 PB of raw storage in a 32-CPU server. This will be used in various use-cases, going from data lake to virtualization.

Creating customer Value with the Next Generation of NUMA

Advantage

The main competitive advantage of a NUMA architecture in general is the fact that local memory accesses are significantly more efficient than remote memory architectures. This is exploited by the OS and applications through information made available to the OS and apps regarding the latency and bandwidth distance between the CPU cores and their respective local memories. The fact that the local memory is much faster than the remote memory and that apps that exploit this can be compared with traditional SMPs where all the memory was equally distant and correspondingly slower. This means that all the memory accesses were suffering from the increasing discrepancy between the speed of the CPU and the more or less constant latency of the DRAM over time. This opening gap in latency has made the modern CPUs much more performance dependent on efficient multi-level cache hierarchies. With multiprocessing NUMA architectures these caches require efficient cache coherency mechanisms to reduce CPU stalls on memory references.

If the coherency mechanisms in the interconnect between the CPUs is inefficient, the CPU efficiency will suffer. Numascale has developed a very efficient directory-based snoop filtering mechanism that reduces the amount of system-wide snoop operations and optimizes the performance of the snoops that cannot be avoided.

The Numascale architecture also supports address interleaving and multiple data-planes with the node controllers in order to provide increased bandwidth and snoop-filter capacity through increasing the number of node controllers used in each system for correspondingly increased performance for mission-critical applications like in-memory database systems.

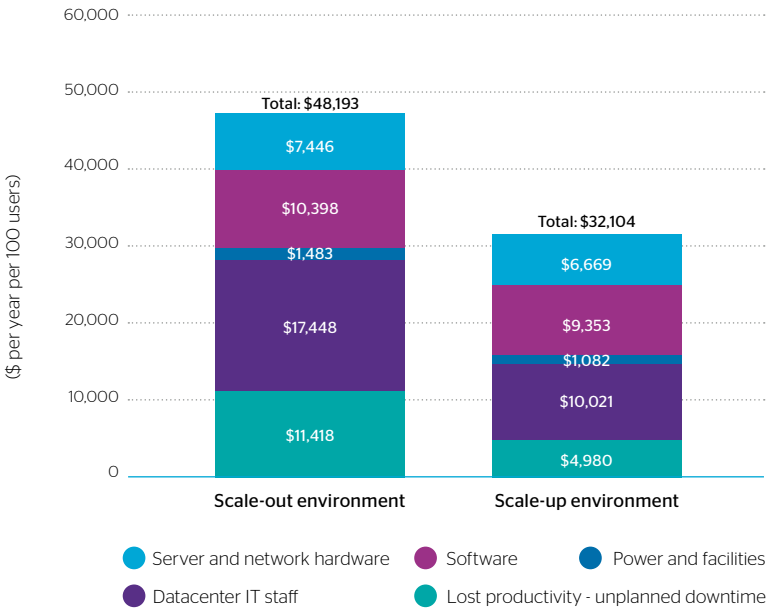
At a high level the Scale-Up provides following advantages over Scale-Out:

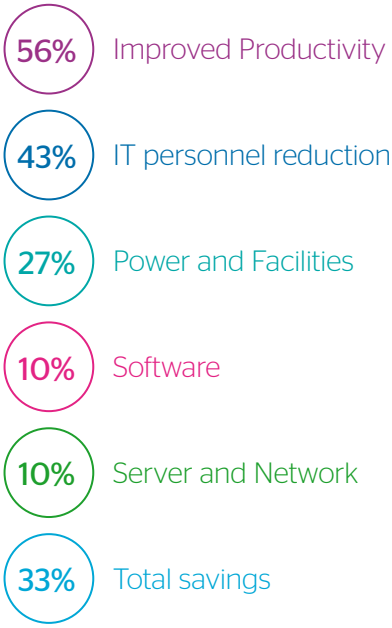
- On-chip multi-dimensional Fabric Interconnect
  Reducing cost and complexity
- Scalable System Size
  Direct and Route-through links
- Linear Cost / Performance scale-up
  TCO friendly
- Modular Architecture
  Customer differentiation
- Micro-coded coherency engines
  Enables microcode patching
- Cache Coherency with Full-Directory based protocol
  Lower Latency and Improved scalability

Annual cost for Scale-Up servers vs. Scale-Out servers in a selected environment for a given number of users, have the following advantages according to IDC:

- 56% Productivity Improvement
- 43% Reduction in IT personnel
- 27% Power and Facilities improvement
- 10% Server and Network improvement
- 10% Software efficiency improvement
- 33% Overall TCO savings

Annual Cost of Scale-Up Servers versus Scale-Out Servers per 100 Users in Selected Environments





Source : IDC Reports

Gen to gen improvement

Based on specific workloads, the future generation of BullSequana S Scale-Up servers will perform up to 40% better than currently deployed BullSequana S servers that are designed on Ultra Path Interconnect (UPI) 1.1. The overall throughput (Bandwidth and Latency) of the server will be greatly improved for large scale in-memory computing applications (e.g. SAP HANA).

The next generation of NUMA architecture, using an external Node Controller (xNC) to interconnect all NUMA nodes of the BullSequana S server, is expected to improve the global latency by more than 30% compared to the current generation. Based on the optimized interconnect topology, the performance improvement is expected to be linear for next generation of BullSequana (10 to 32 sockets) Scale-Up servers.

Intel processor with higher number of UPI links (up to 4) enables next generation of BullSequana servers to effectively take advantage of NUMA awareness of specific applications with 4 NUMA nodes concurrently.

Next generation of BullSequana S server will maintain 2-socket upgrade modularity, similar to current generation. This will enable end-users to build higher throughput Scale-Up configuration with more processors (2 sockets at a time) based on the usage demand minimizing the extra cost for unneeded CPU and memory. This flexible architecture is critical to achieve cost optimized investment for running many business applications.

The next generation xNC is designed to fully support the next generation of the UPI protocol with all speeds, bringing bandwidth on UPI links up to 20+ GT/s. The bandwidth of the UPI protocol, used on both glueless and xNC configurations, will be more than 100% improvement.

In addition, the memory bandwidth will also be more than double, contributed by more memory channels (from 6 to 8 per CPU) and higher speed DDR5 compared to DDR4 DIMMs (from 2666MT/s to 4400MT/s).

The roadmap of BullSequana S Scale-Up servers is in cadence with Intel Server CPU roadmap keeping forward compatibility as supported by UPI technology from one generation to the next. The next generation of BullSequana S will support large memory footprint up to 192TB with 32-Socket (using 256GB DDR5 and 512GB Optane) for next generation of Intel server platform. New xNC design may be extended to support 640TB with 32-Socket system.

Global packaging of the server will allow to fit in a standard rack a 16s-configuration by using 19 Rack Units. A 32-socket configuration will use 38 Rack Units and can still fit in a standard 42RU rack.

Current Gen 32-socket BullSequana S (model S3200)

Scale-Up Future

Scale-Up computing is neither a new concept nor a new computing need. The rise of volumetric data drove the perpetual need of high throughput NUMA SMP class compute capability.

This has been evolved over many years from 1S, to 2S, to 4S, 8S, 16S, and 32S on X86 architecture platform. External node controller is required to go beyond 8S on x86 server platform.

This is where Intel, Atos and Numascale are playing a key role effectively enabling scale-up compute ecosystem.

Atos and Numascale are well poised to address Scale-Up computing with BullSequana S class server generation over generation addressing mission critical workloads (Big data, AI, IoT) with reduced Total Cost of Ownership (TCO).

Authors

Atos: Sylvie Lesmanne, Paul Magadalena, Pascale Martinez, Alexandra Roy, Didier Marcon

Numascale: Goutam Debnath, Einar Rustad

Intel: Jake Smith

About Atos

Atos is a global leader in digital transformation with 110,000 employees and annual revenue of € 12 billion.

European number one in cybersecurity, cloud and high performance computing, the group provides tailored end-to-end solutions for all industries in 73 countries. A pioneer in decarbonization services and products, Atos is committed to a secure and decarbonized digital for its clients. Atos operates under the brands Atos and Atos|Syntel. Atos is a SE (Societas Europaea), listed on the CAC40 Paris stock index.

The purpose of Atos is to help design the future of the information space. Its expertise and services support the development of knowledge, education and research in a multicultural approach and contribute to the development of scientific and technological excellence. Across the world, the group enables its customers and employees, and members of societies at large to live, work and develop sustainably, in a safe and secure information space.

Find out more about us
atos.net
atos.net/careers

Let's start a discussion together

About Numascale

Numascale® is a fabless semiconductor company and the leading provider of Cache-Coherent shared memory interconnect technology for modern CPUs.

Numascale® developed Intel® UPI 1.0 based Cc-Numa Node Controller for Intel® Xeon® Scalable Processors, supporting scale-up servers up to 32 Xeon® CPUs and can support up to 48TB shared memory with 128GB DIMMs using single instance of operating system.

The Scale-up servers are well suited for big data, artificial intelligence, and internet of things. These servers have high RAS features and up to 33% TCO savings.

Server systems using Numascale® Node Controllers are marketed, sold, and supported world-wide by Atos and its partners.

More about Numascale can be found at:
Numascale.com

Let's start a discussion together

CT-210203-RY-J1688-ATOS-WP-RISEOFVOLUM