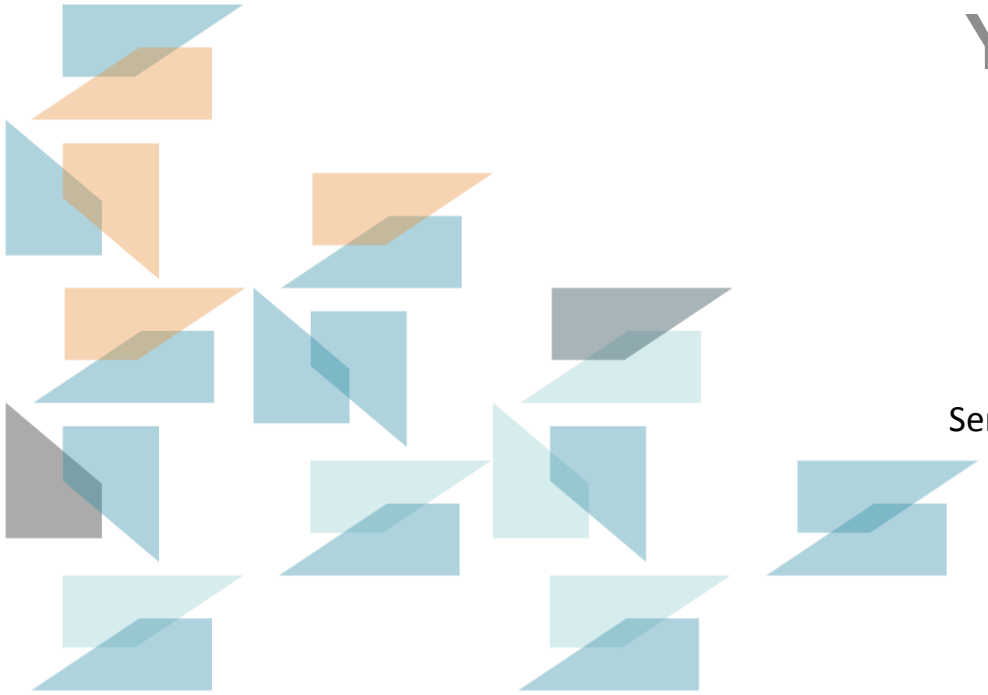
Simulating large datasets using 2,3 TB of memory, 576 cores and 12 GPUs in one single Linux system

Scalable Solutions Designed to Grow with Your Needs



Atle Vesterkjær

Senior Software Engineer

av@numascale.com

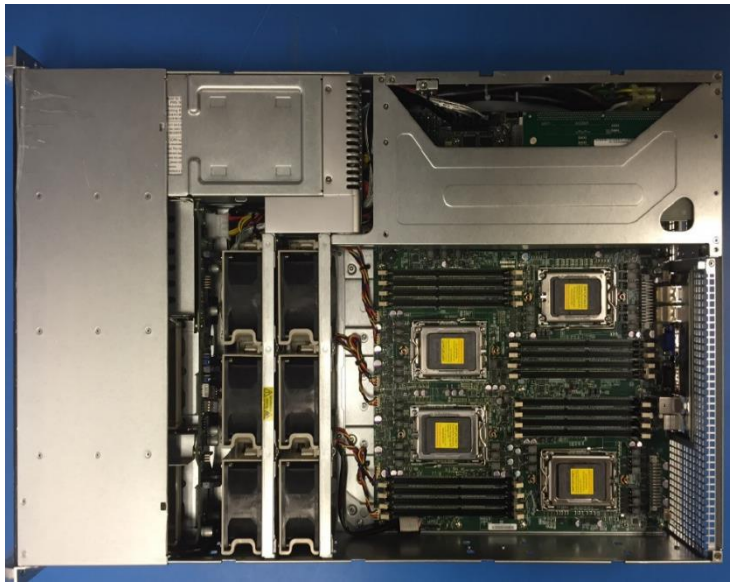
February 2016

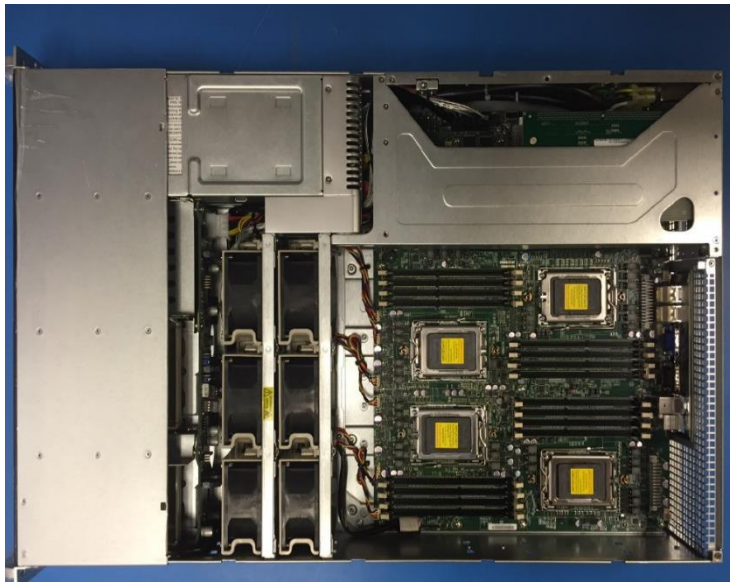
NUMASCALE
BIGGER DATA ANALYTICS

Need more cpus?

Traditional clustering solutions has unatural boundaries in:

- RAM per server
- CPUs per server

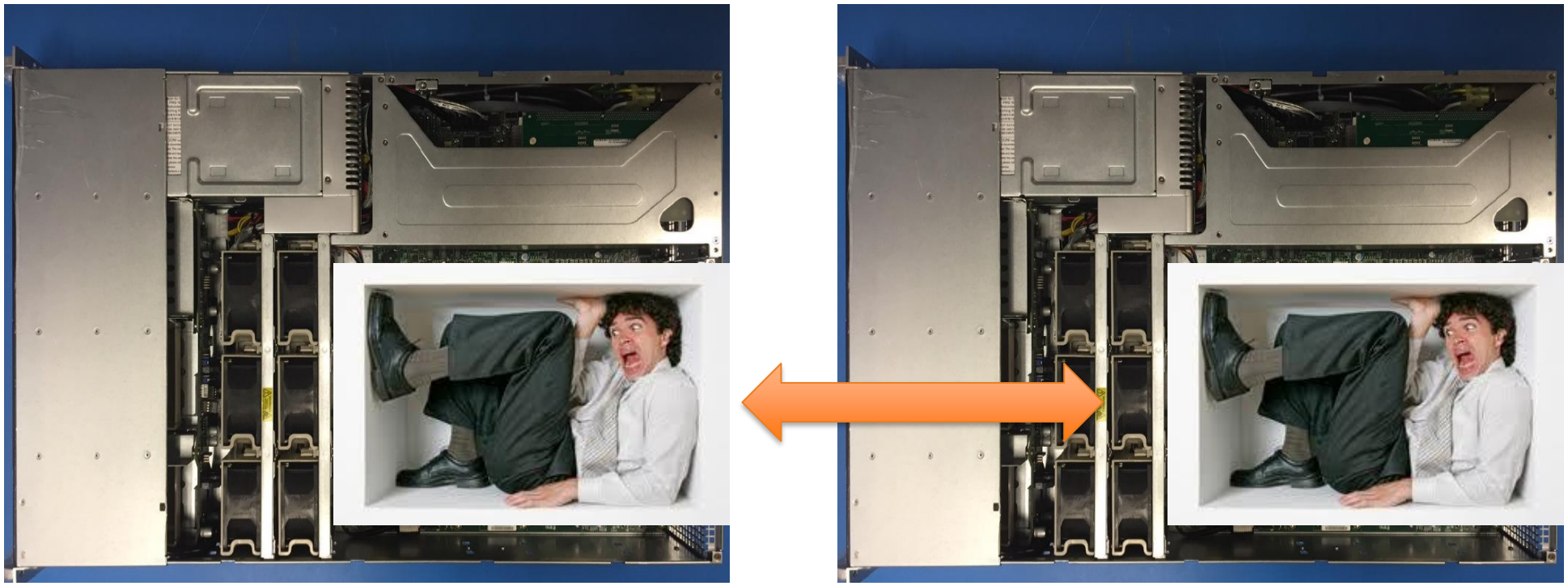




Out of memory space?

Traditional clustering solutions combine servers using

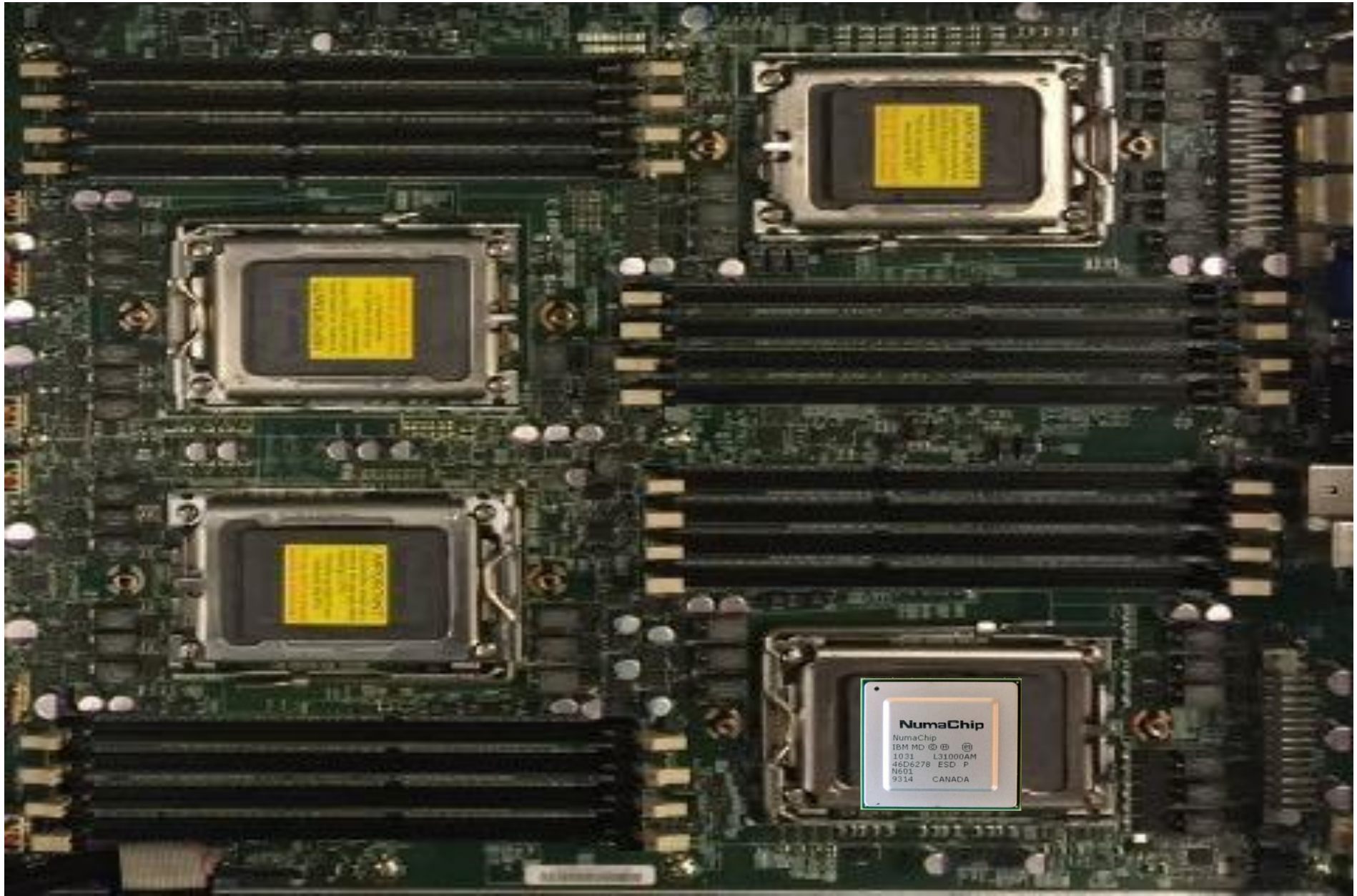
- Complex programming models
- IO adapters



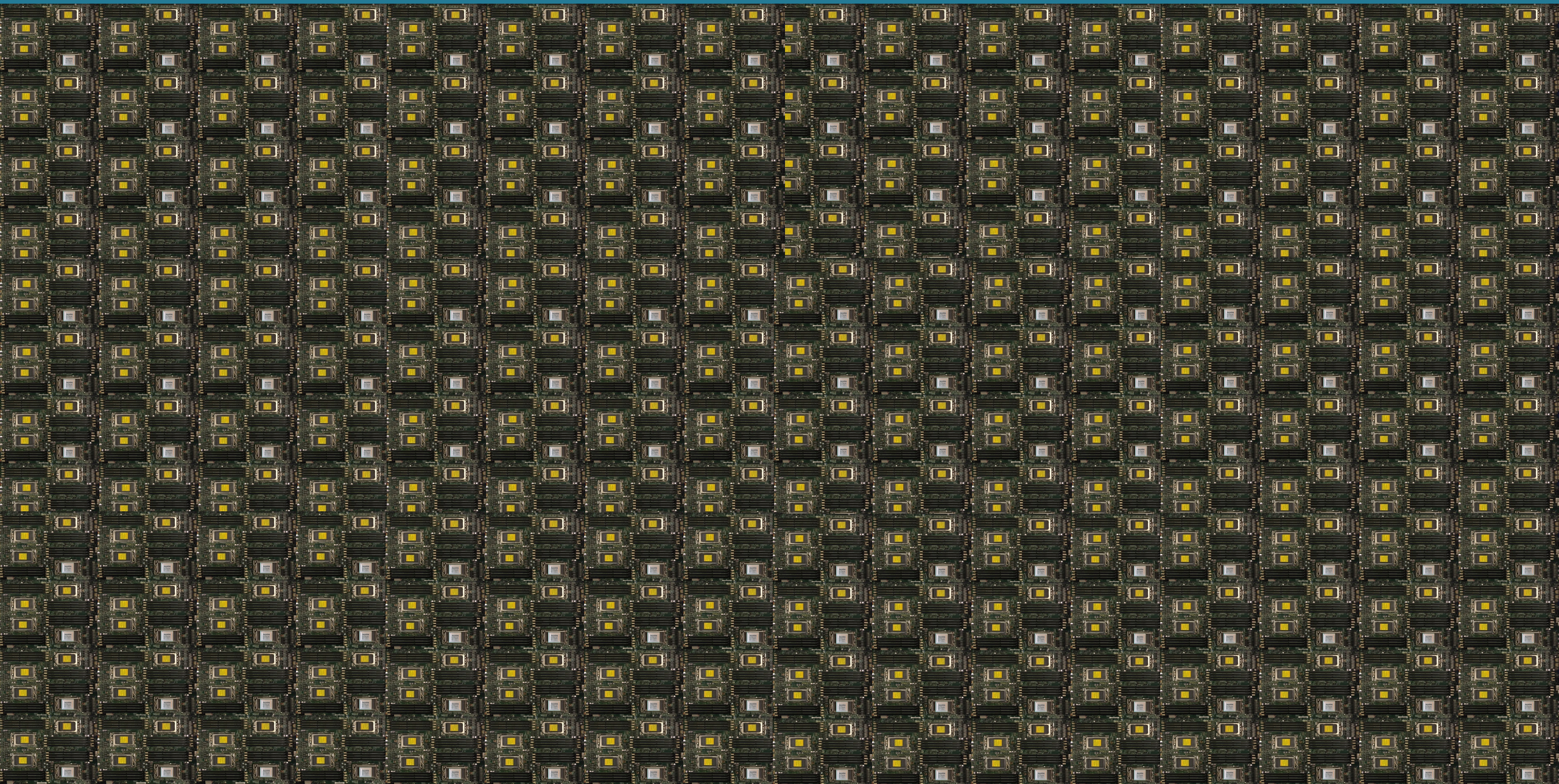
Your system can grow organically with Numascale

Enourmous resources has been used to adapt to unatural boundaries

It does not have to be this way



Using the Native CPU communication protocol

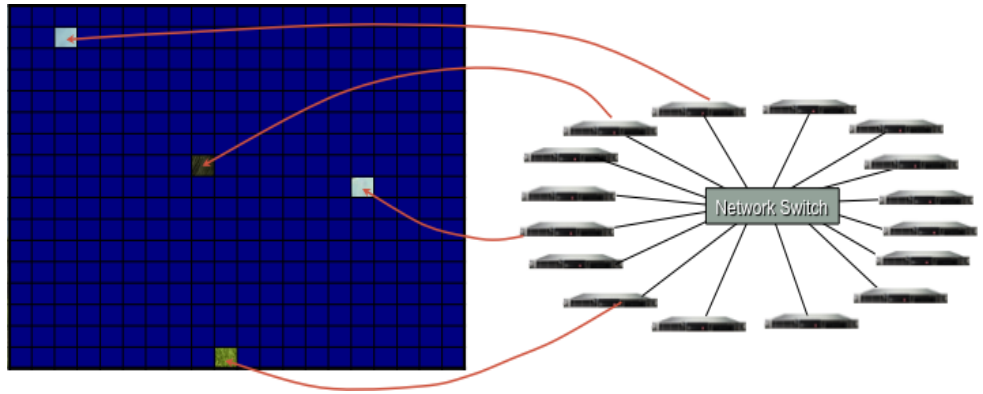


Numascale's Unique Solution

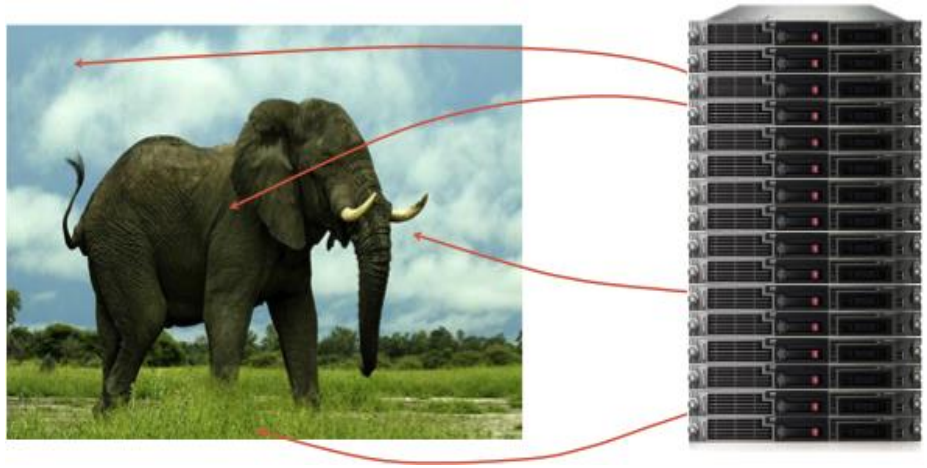
Numascale enables:

1. IN-MEMORY computing of large data sets at 100X higher speed by reducing interconnect latencies and avoiding disk paging = Efficient Predictive Analytics.
2. Powerful IO allows for real-time sensor and data input and rapid time to insights = Visualization.
3. Simplified software development, system utilization and reduced adm. and maintenance overhead by running single OS.
4. Low entry level cost and unmatched scalability

Traditional Cluster:
Distributed memory gives partial views of data set



Numascale "cluster":
Shared memory gives shared view of entire data set



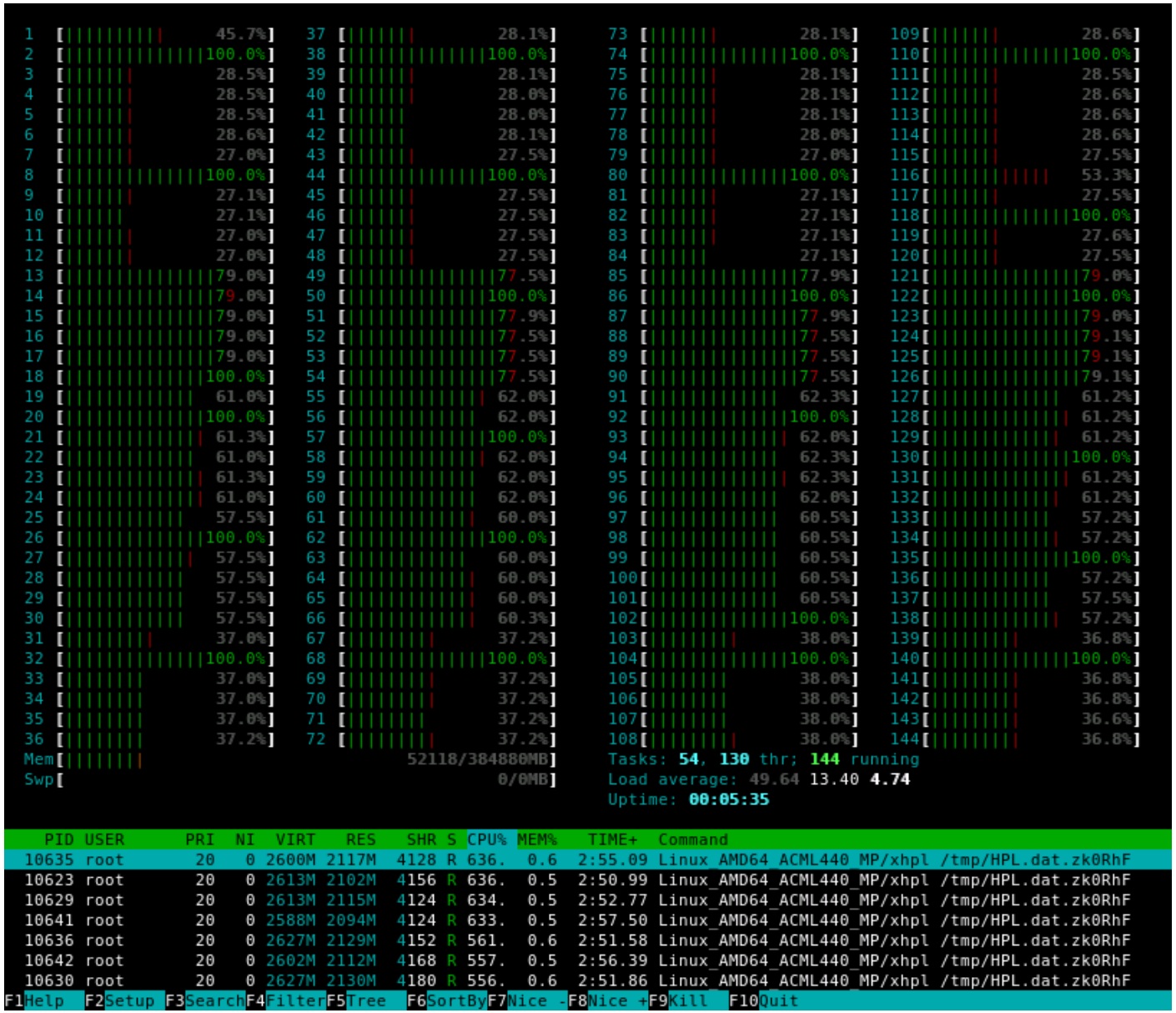
Unbeatable Programming Model

- Single System Image
- All applications and APIs
- NO Virtualization SW Layer!
- One System

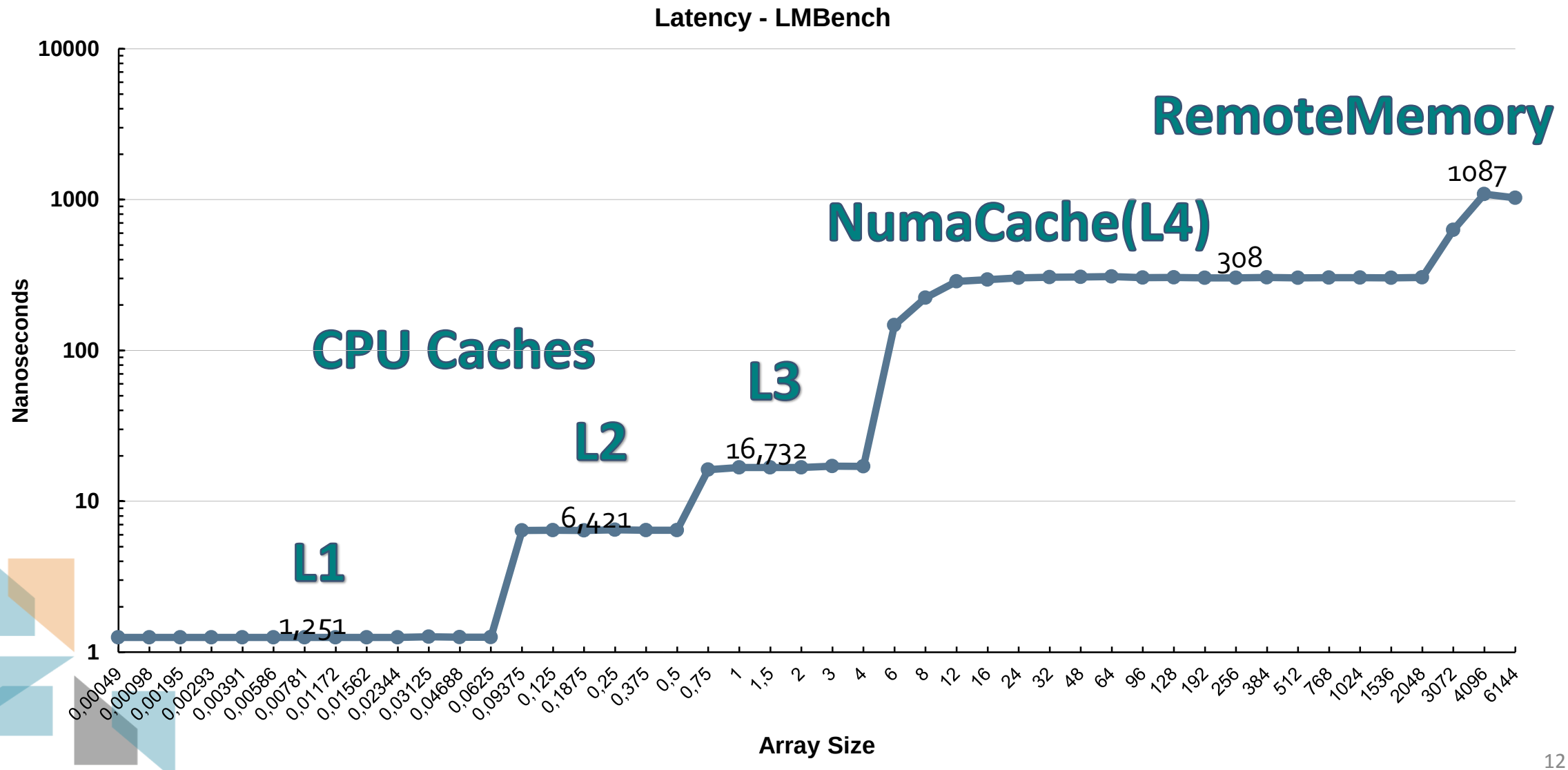
```
top - 09:29:57 up 20 days, 20:07,  6 users,  load average: 8.32, 8.31, 8.32
Tasks: 3068 total,   3 running, 3065 sleeping,   0 stopped,   0 zombie
Cpu(s):  1.4%us,  0.1%sy,  0.0%ni, 98.5%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  1583148160k total, 1185082348k used, 398065812k free,        4k buffers
Swap: 33046524k total,        0k used, 33046524k free,  4039576k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
202867 root      20   0 1119g 1.1t  524 R  815 74.1   3758:34 x.mod2as
197024 root      20   0 19160 6560 1204 R   37  0.0 403:40.55 htop
206397 root      20   0 13512 3704  924 R   14  0.0   0:04.48 top
197023 root      20   0 13512 3700  924 S   13  0.0 177:20.24 top
  1399 root      20   0     0    0    0 S    3  0.0   1:03.79 ksoftirqd/348
  2187 root      20   0     0    0    0 S    3  0.0   0:50.98 ksoftirqd/545
    10 root      20   0     0    0    0 S    2  0.0 290:26.05 rcu_sched
 11190 root      20   0     0    0    0 S    1  0.0  68:02.65 kworker/40:1
 11177 root      20   0     0    0    0 S    1  0.0 173:47.85 kworker/24:1
 11241 root      20   0     0    0    0 S    1  0.0  86:23.99 kworker/44:1
 11668 root      20   0     0    0    0 S    1  0.0  53:37.71 kworker/348:1
 11688 root      20   0     0    0    0 S    1  0.0  59:31.14 kworker/336:1
 12035 root      20   0     0    0    0 S    1  0.0  20:37.15 kworker/545:1
 11197 root      20   0     0    0    0 S    0  0.0  96:09.80 kworker/36:1
 11203 root      20   0     0    0    0 S    0  0.0 148:15.71 kworker/32:1
 11209 root      20   0     0    0    0 S    0  0.0 166:31.05 kworker/28:1
 11233 root      20   0     0    0    0 S    0  0.0  82:54.52 kworker/48:1
 11626 root      20   0     0    0    0 S    0  0.0  32:13.59 kworker/308:1
 11710 root      20   0     0    0    0 S    0  0.0  26:04.65 kworker/357:1
 11714 root      20   0     0    0    0 S    0  0.0  31:09.39 kworker/354:1
 47717 root      20   0  7764  576  484 S    0  0.0  22:06.24 tail
 47736 root      20   0  7764  576  484 S    0  0.0  22:13.57 tail
```

Natural evolution of previous breakthroughs

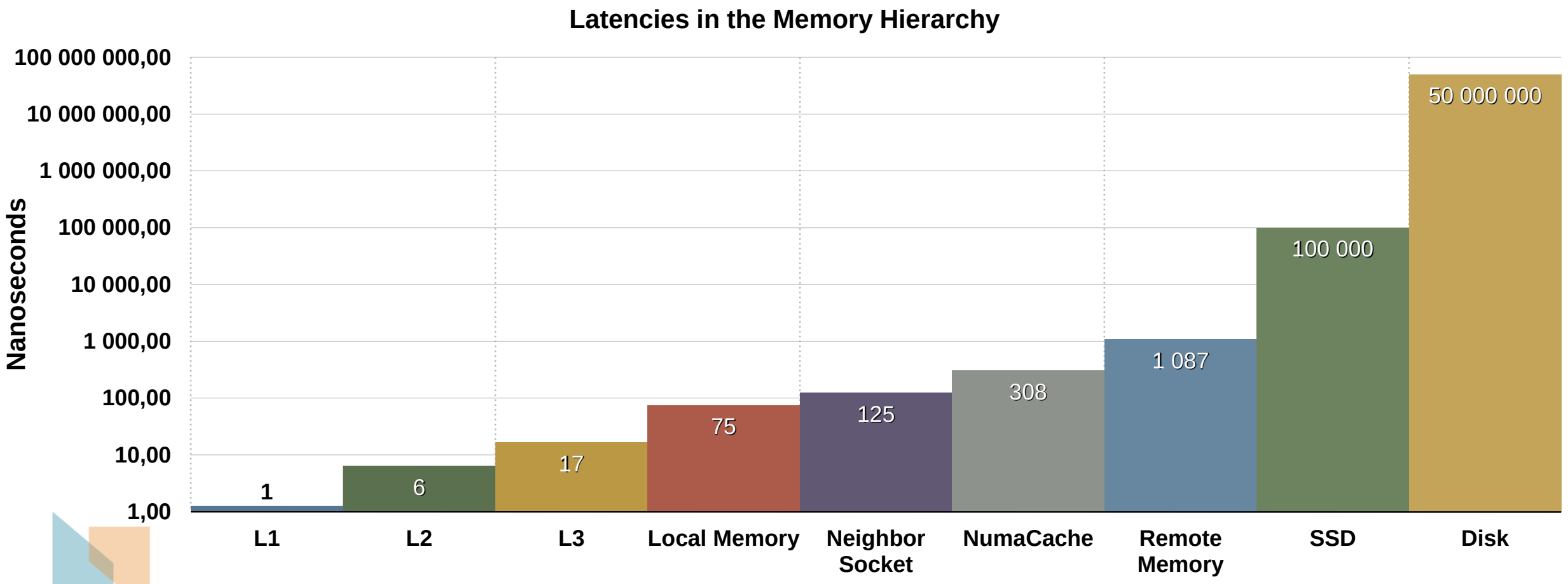
- Step 1: Singlecore PC
- Step 2: Multicore PC
- Step 3: Multisocket Server
- Step 4: Scale out with bus bridging interconnect architechture
- Finally: Numascale systems share GPUs, CPUs, memory and IO between physical servers



LMbench - Chart

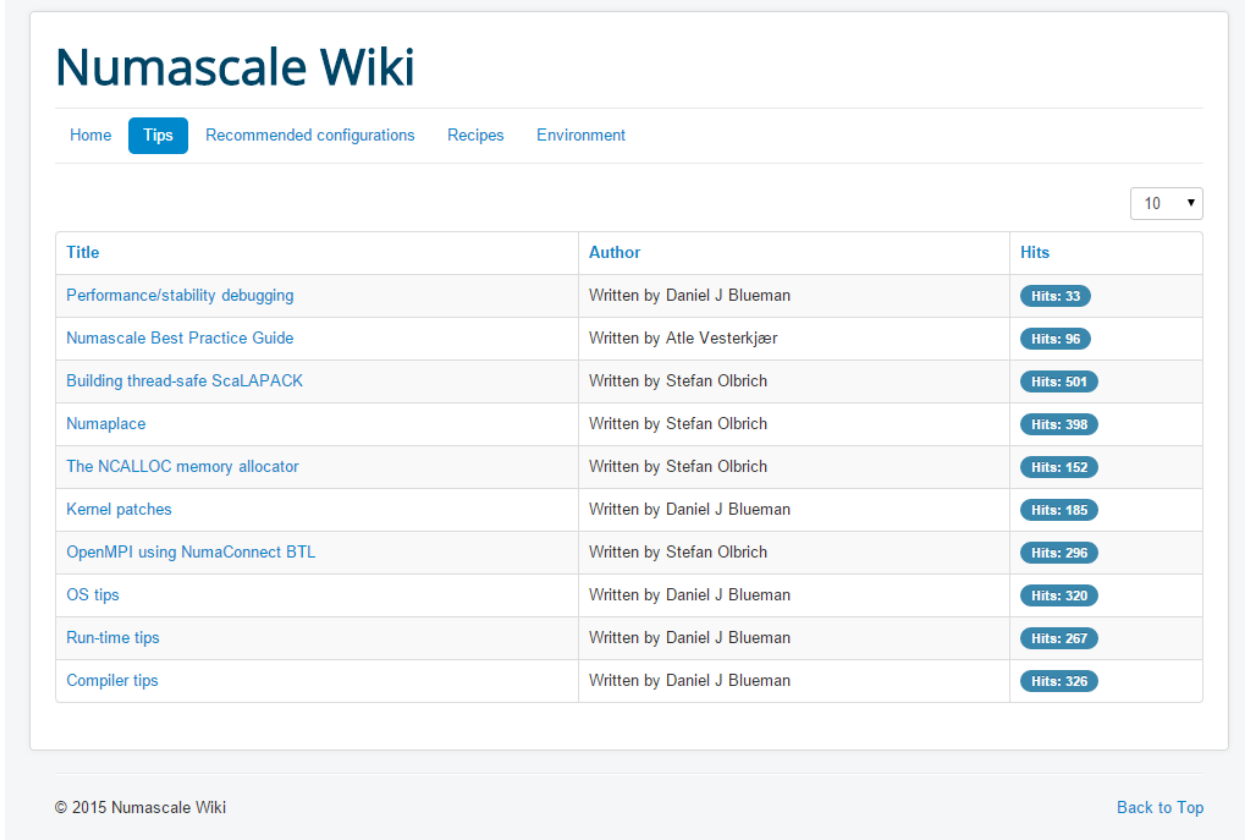


The Memory Hierarchy



Braking records with Open Source Software

- Numascale Appliances comes with Open Source Software Stack.
- Numascale builds larger single image systems than anyone else.
- Patches to the Linux kernel, libraries and tools returned to the open source community



Easier Parallel Programming

- Any **multithreaded** application can run on NumaConnect using e.g pthreads.
- **OpenMP** can be used for easily applying a parallel programming model for a shared memory computer and will scale the execution of parallell regions in an application across many threads
- Numascale develops preloaded modules for R, **OpenMPI, LAPACK, memory allocation and more** that fully utilize the NumaConnect

Worlds Largest SMP

108 nodes
324 AMD 6386 CPUs
5184 cores
20.7TBytes Memory

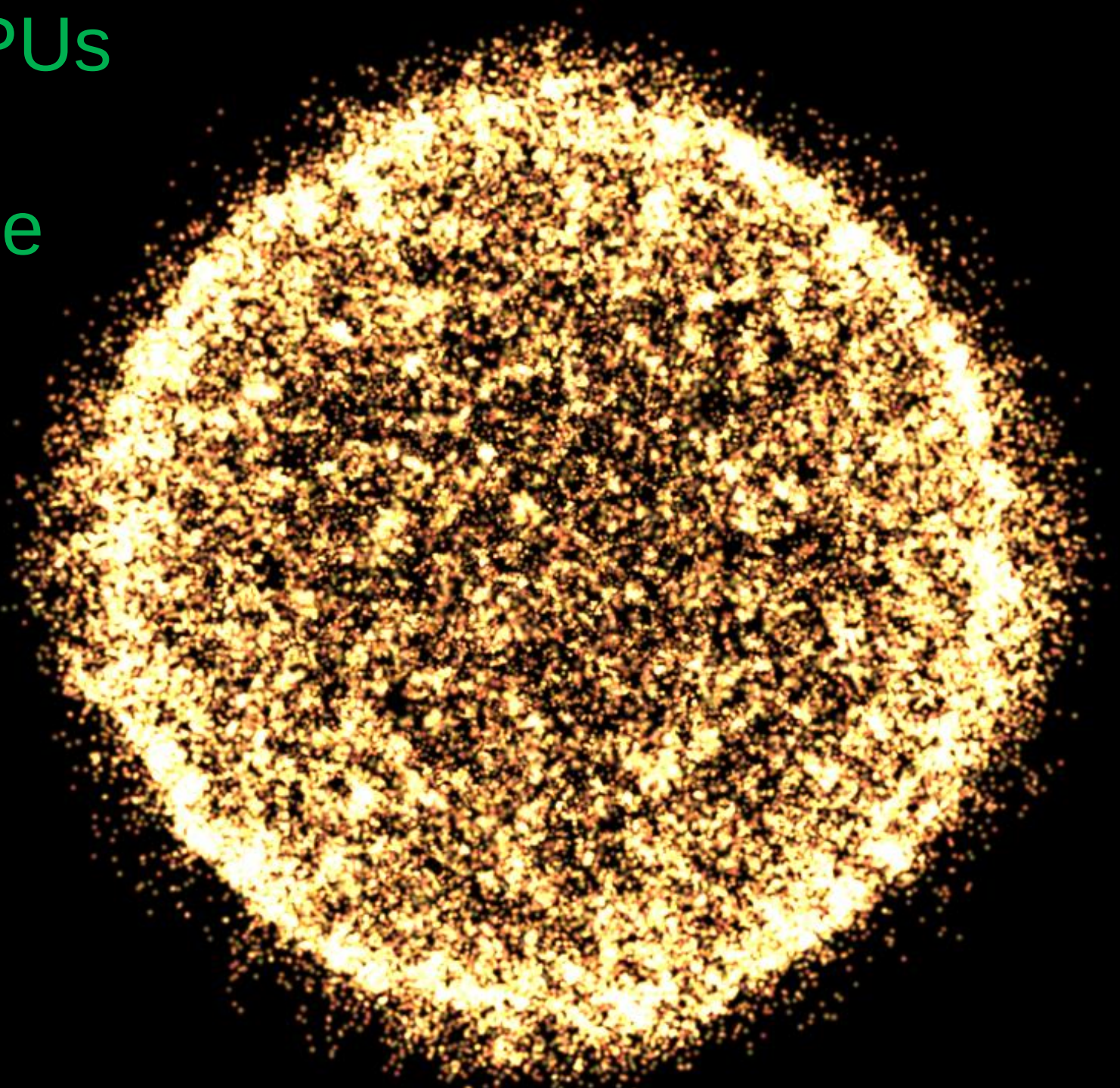
Application: Real time Big Data Analytics

World Record Memory Bandwidth

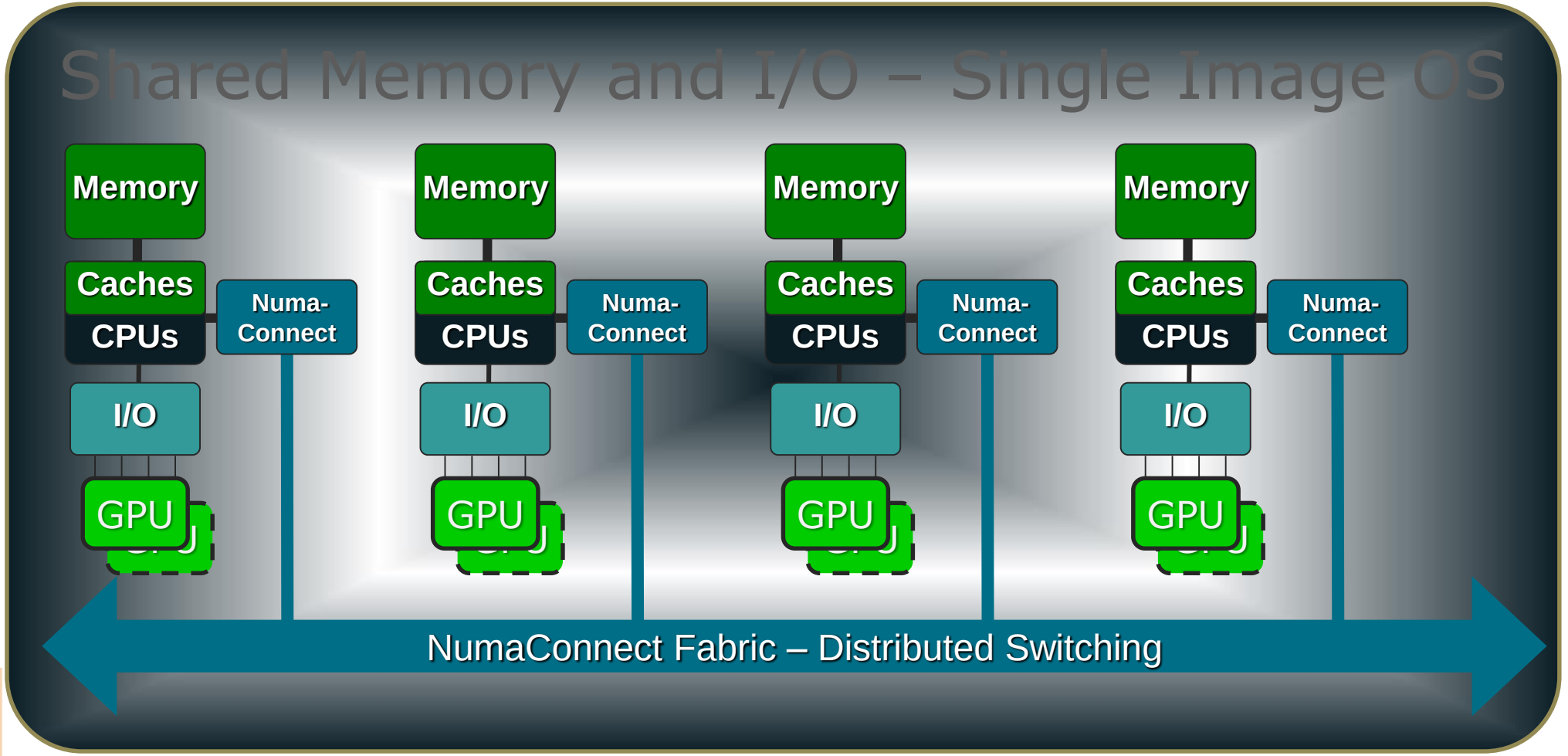
```
-------------------------------------------------------------
STREAM version $Revision: 5.10 $
-------------------------------------------------------------
This system uses 8 bytes per array element.
-------------------------------------------------------------
Array size = 135000000000 (elements), Offset = 0
Total memory required = 3089904.8 MiB (= 3017.5 GiB).
Each test is run 10 times, but only
the *best* time for each is used.
-------------------------------------------------------------
Number of Threads requested = 1296
-------------------------------------------------------------
Function    Best Rate MB/s  Avg time     Min time     Max time
Copy:        9139226.2     0.240336     0.236344     0.246761
Scale:      10062237.6     0.217982     0.214664     0.220526
Add:         8985643.0     0.361473     0.360575     0.363363
Triad:       8871850.0     0.366032     0.365200     0.366646
```

Human readable numbers:
Array size = 3 TB
Copy = 9.1 TB/s
Scale =10,0 TB/s
Add = 9,0 TB/s
Triad = 8.9 TB/s

Scalable GPUs with Numascale



System Architecture



GPU Computing

- Standard GPU Clusters suffer from
  - Complex message based inter-node communication
  - Limited node size
  - Each node requires one instance of the OS
    - Complicates system administration
    - Complicates aggregation of GPUs
  - Underutilized CPUs

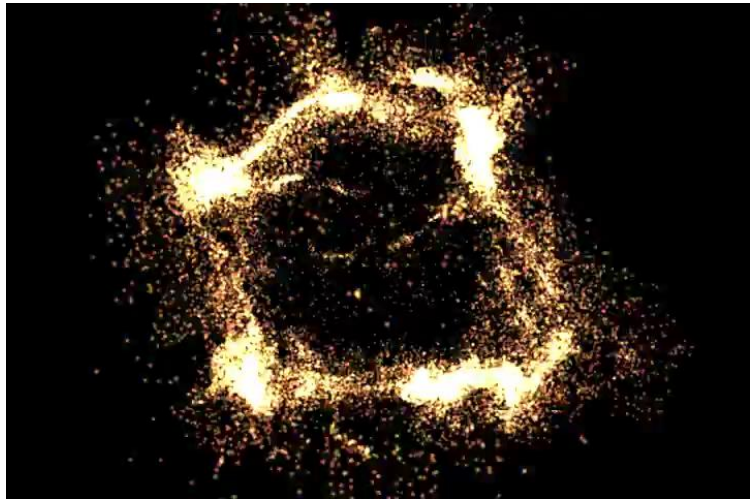
Visualization in Clouds

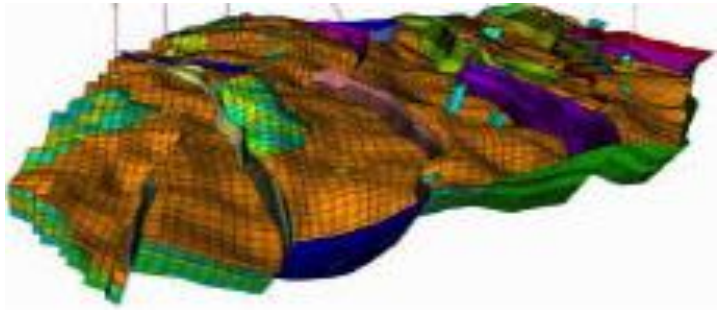
- Sharing GPU Resources
  - Fewer and larger systems
  - Better Resource Utilization
  - Less system administration
  - Easier upgrades

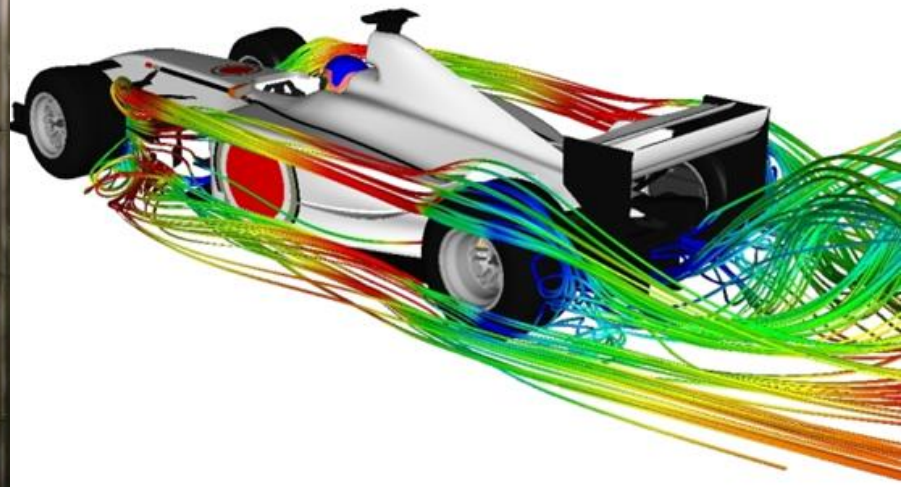
Scalable GPUs

Supermicro 2042 Server Features:

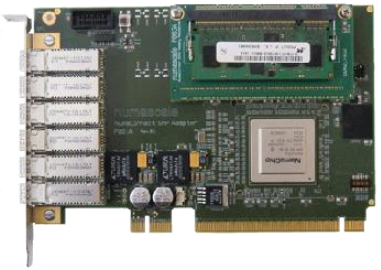
- Supermicro 2042G-LTF Server
- 2U Rack Mount
- 1 GPU - Single/Double Slot
- 2-3 AMD 63xx CPU
- 8 to 48 cores
- Up to 3.5GHz
- Up to 384 GB DRAM
- NumaConnect Adapter
- 1400W Redundant High-efficiency (Gold) Power Supplies
- Up to 6 hot-swap SAS/SATA



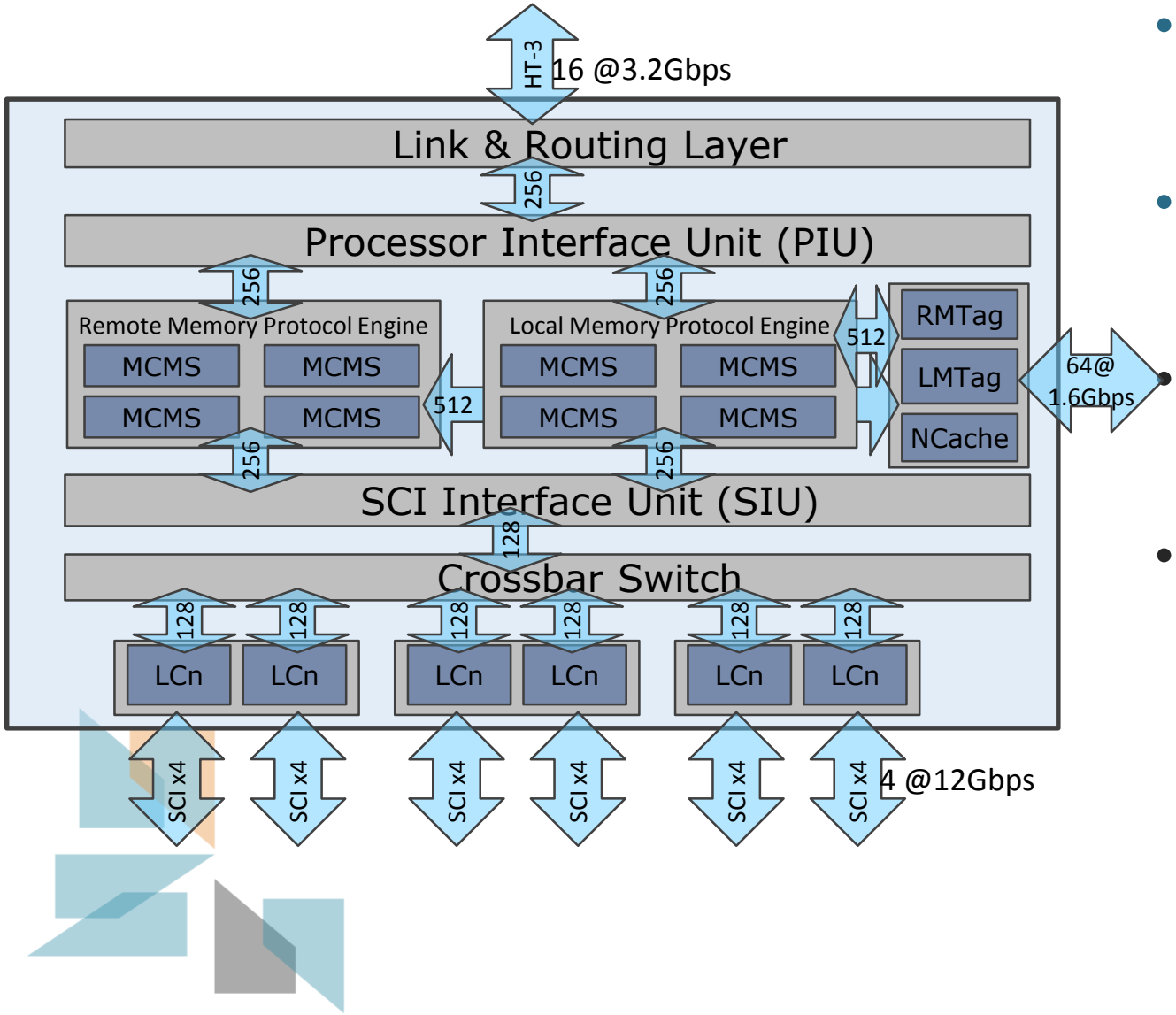




4 GPUs single image

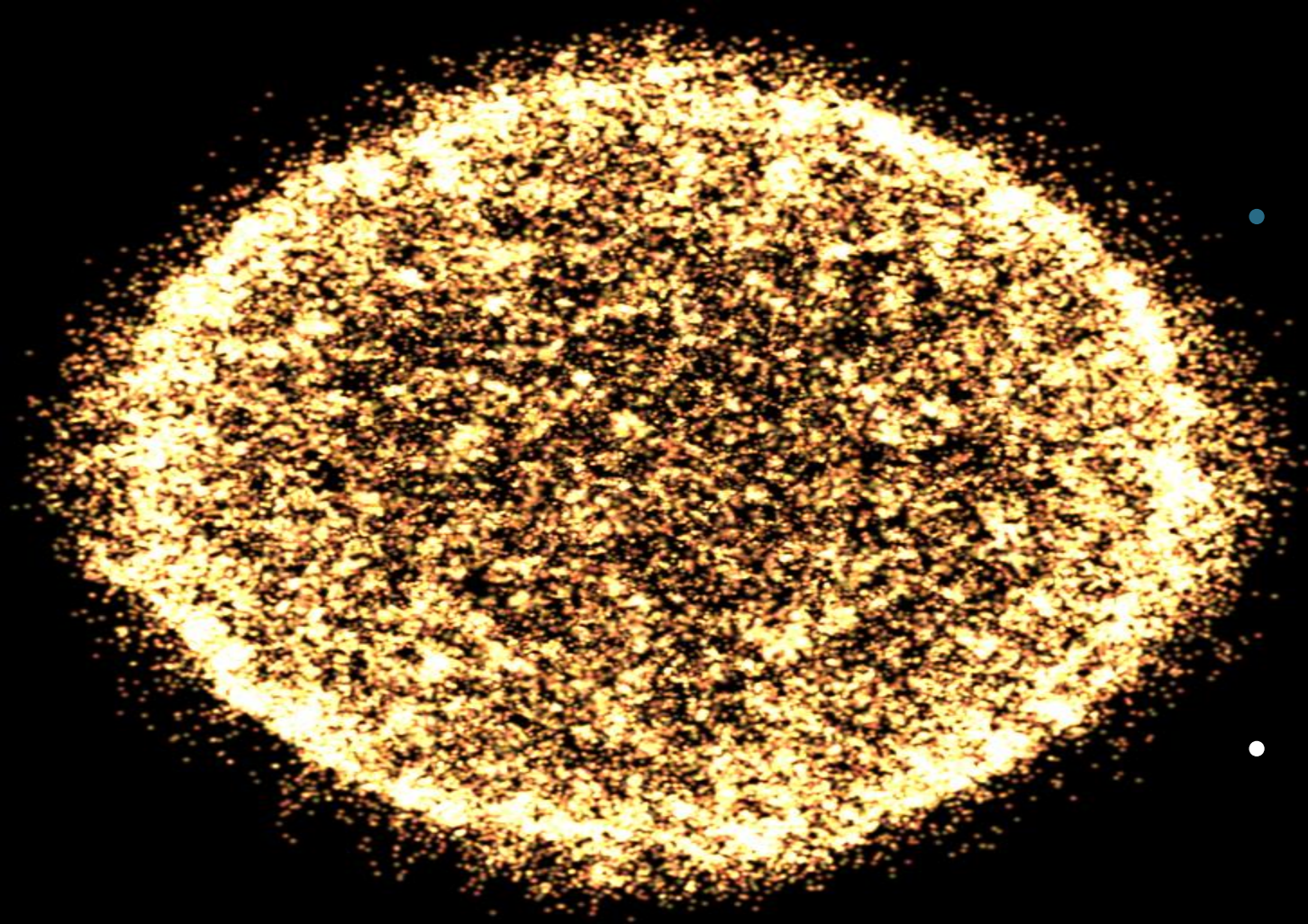


News



- **NumaChip2** released with fatter pipes which gives more paralellism and BTE engine, both gives **more bandwidth**
- Numascale R Appliance released with optimized software stack and plug and play experience for Data Scientists
- Numascale has developed their own simple Jacobi solver that shows linear scaling for the Shyne code.
- Numascale is developing their own solver,
  for **solving** large **sparse** symmetric and unsymmetrical linear systems of equations on large shared-memory multiprocessors by maximum utilization of **NumaChip2.**

Future system at Keele



24.36 FPS [single precision | 65536 bodies]
2092.64 GFLOP/s on 3 GPUs

- **One Numascale system**
  - 2,3 TB of memory
  - 576 cores
  - 12 GPUs.
- All recourses are available in one Linux instance

Share everything

Thank you

Atle Vesterkjær

Senior Software Engineer

av@numascale.com

February 2016