OpenFOAM Scaling on Cray Supercomputers

Dr. Stephen Sachs | GOFUN 2017

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Supercomputing Leadership

Since Its Founding, Cray Has Maintained a Single Focus on Supercomputing

Supercomputing Leadership

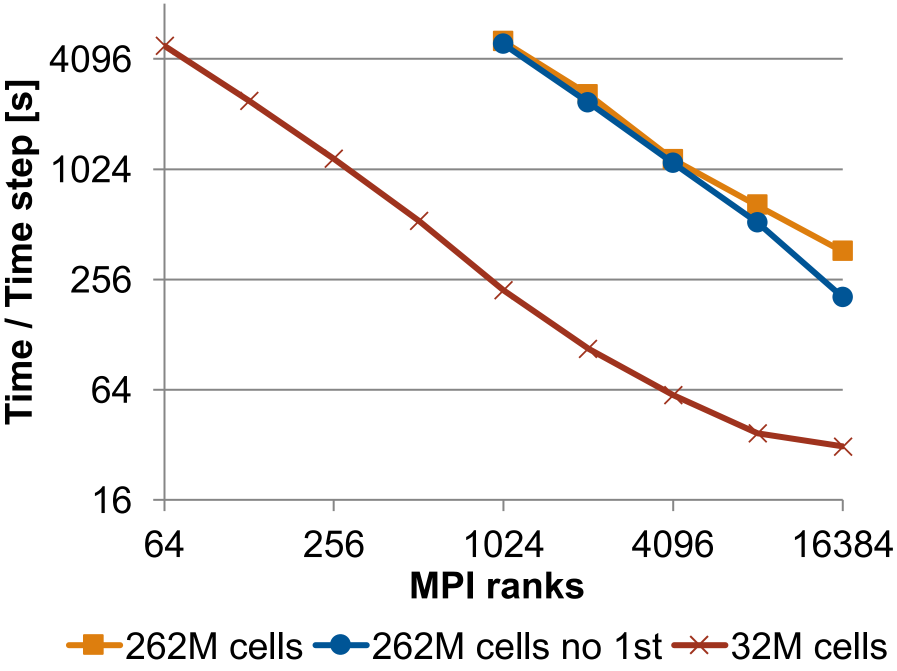
OpenFOAM a HPC code?

- **Portability to HPC architectures – Yes**
  - Support for multiple compilers
  - Support for Intel MIC, GPU
- **MPI parallelism – Yes**
  - Scalability limit could be improved
- **Hybrid parallelism – Partial**
  - Multiple attempts
  - Not in main release
- **High Performance I/O – No**
  - Design follows structure contradicting HPC parallel file system

There is potential | Old slide from 2012

- Inflated cavity tutorial
- AMD Opteron 6276
- OpenFOAM/2.2.0
- icoFoam

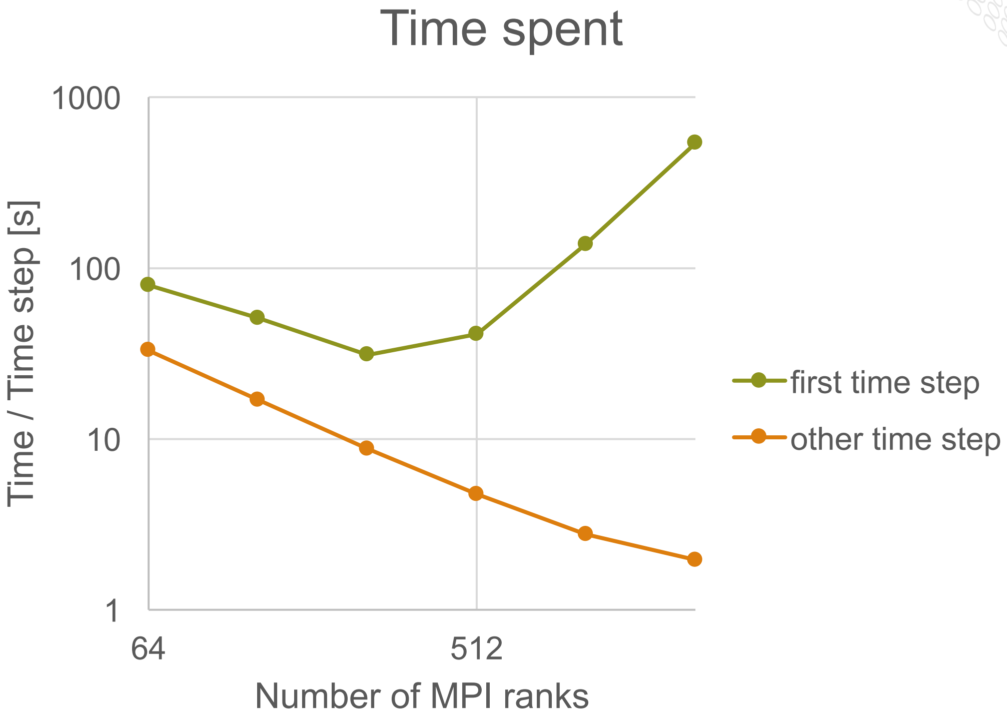
1.0 PFLOP/s
3,552 nodes
AMD Interlagos 32 cores/node



HPC Programmers Wish List for OpenFOAM

Some wishes have been granted | Startup phase

- Communication pattern in initialization
- Serial I/O reading one file per MPI rank
- First time step often skipped in benchmarks



Comparison at Scale

Laptop

- 4 cores
- 2 Mem Channels
- 1 Disk
- 1 Metadata Target

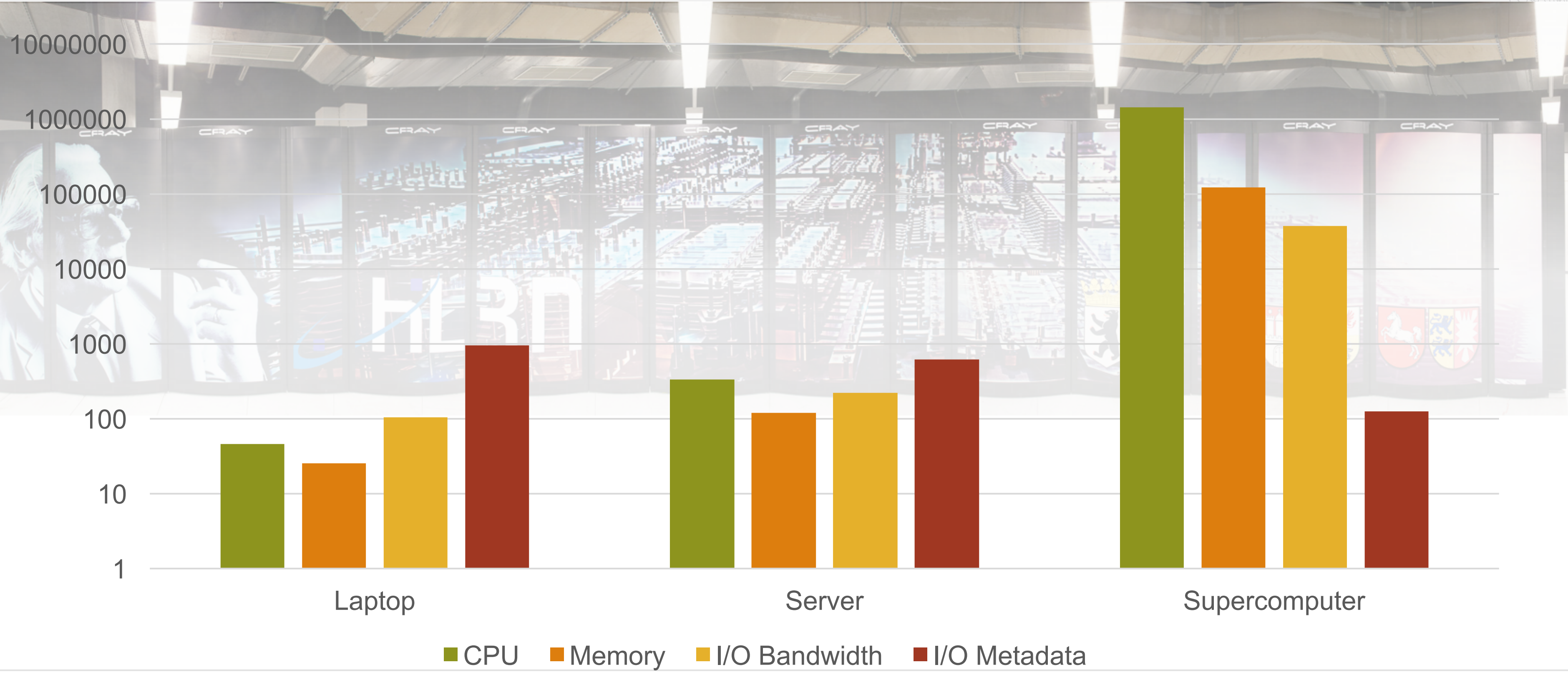
Server

- 20 cores
- 8 Mem Channels
- 6 Disks
- 1 Metadata Target

Comparison at Scale

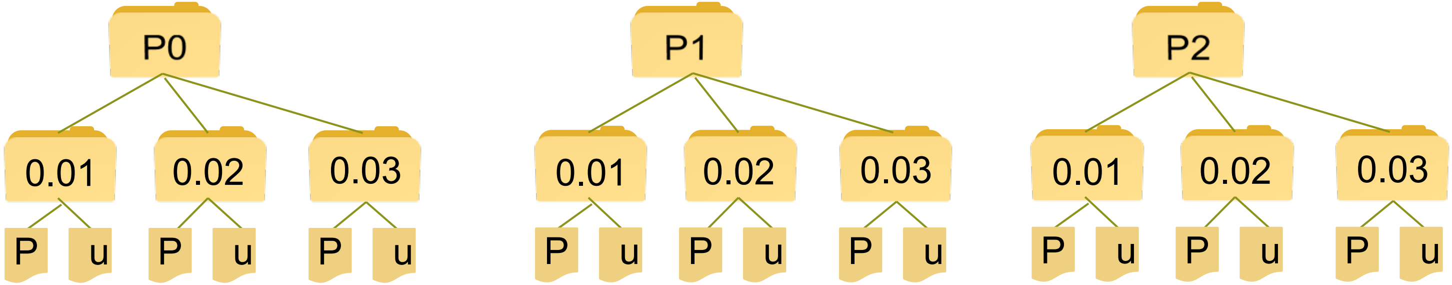
- 44,928 cores
- 14,976 Mem Channels
- 1480 Disks
- 1 Metadata Server

Comparison at Scale



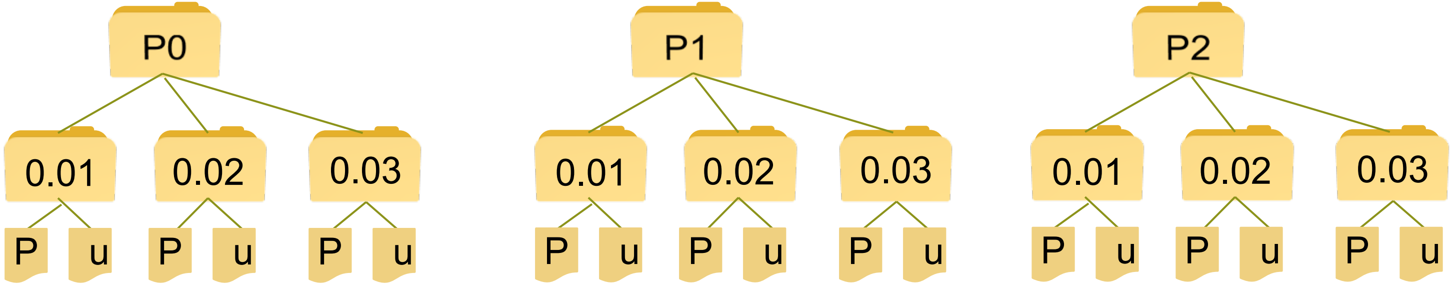
Metadata Operations

- Every time a file is opened or checked
  - Files in the respective directory and subdirectories are checked
- Workstation
  - 1-8 MPI ranks are doing 60-480 metadata calls per second – OK



Metadata Operations

- Every time a file is opened or checked
  - Files in the respective directory and subdirectories are checked
- Workstation
  - 1-8 MPI ranks are doing 60-480 metadata calls per second – OK
- Supercomputer
  - 1.000-10.000 MPI ranks are doing 60.000-600.000 metadata calls per second - Problem



What is in our Toolbox?

Optimization at Scale

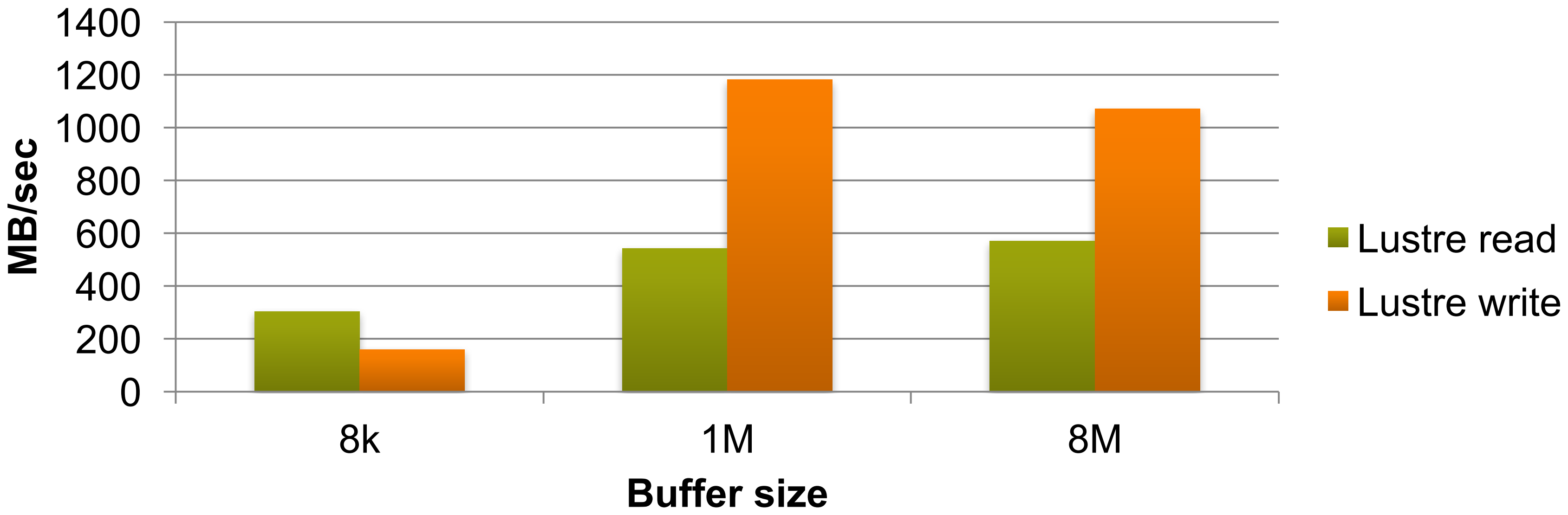
- **Inspect solvers at scale**
  - In case of strong scaling issues
  - GAMG runs faster than PCG, but scales worse
- **Do not check for file changes**
  - Disable runTimeModifiable
  - Accessing metadata can be a source for congestion
- **More MPI messages on the Eager 0 path**
  - Valid for MPICH derivatives
  - Saves one copy for most messages

Optimization at Scale (cont.)

- **Use Huge pages**
  - Larger memory pages can increase memory performance
- **Underpopulate compute Nodes**
  - OpenFOAM is memory bandwidth sensitive
- **Decomposition is key**
  - Scalability limit due to load imbalance
- **Hardware Collection Engine**
  - Offload MPI work to NIC

Optimization at Scale (cont.)

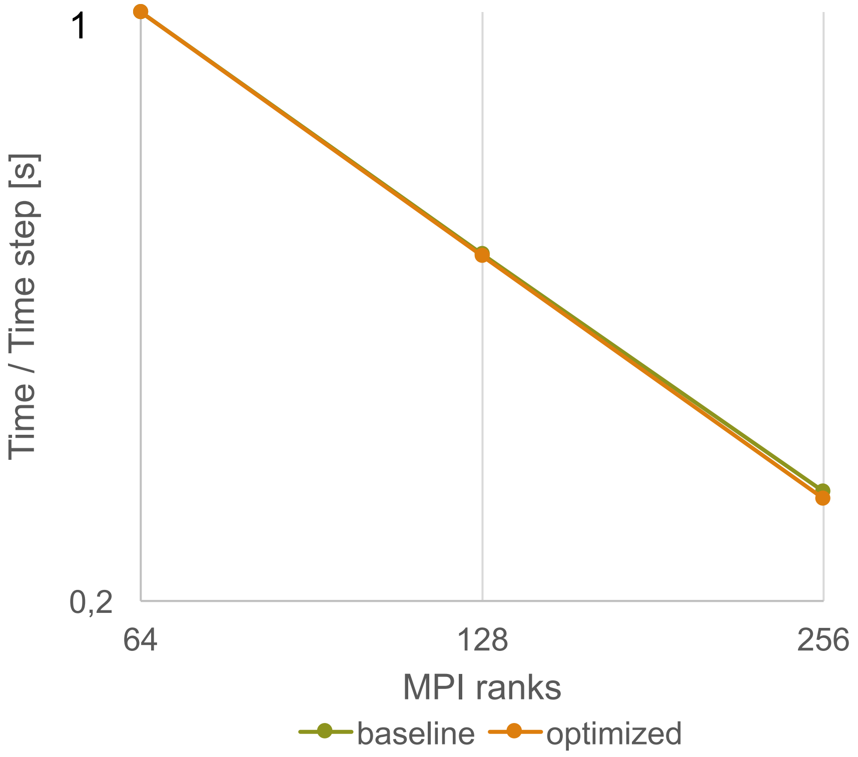
- Standard buffer size does not take advantage of high bandwidth file system



Scalability Results

- OpenFOAM/2.2.2
- ~100M cells
- GAMG solver
- Intel E5-2698 v3 @ 2.30GHz

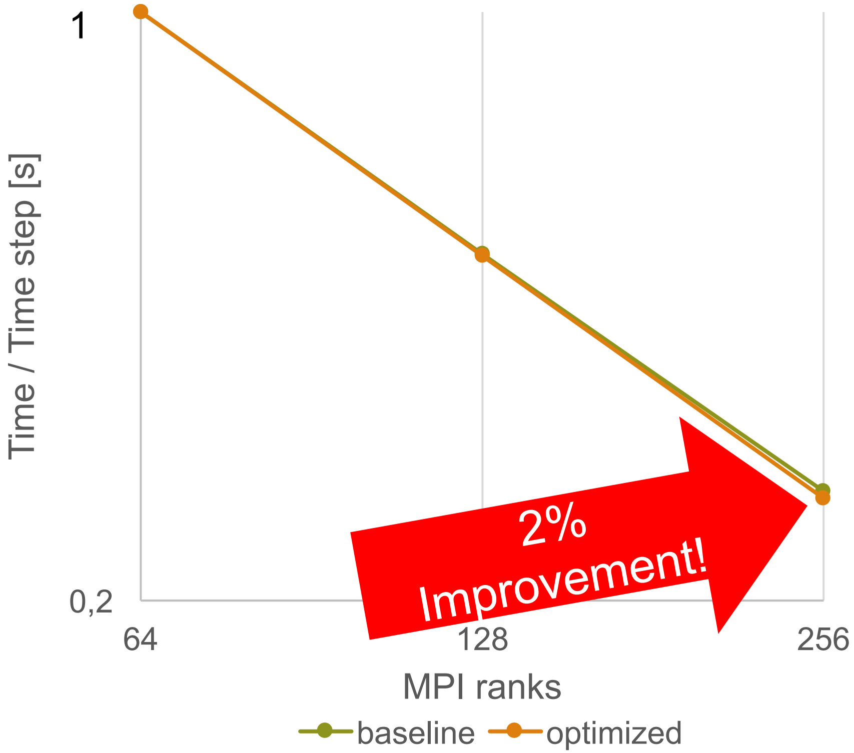
- 3 weeks of work...



Scalability Results

- OpenFOAM/2.2.2
- ~100M cells
- GAMG solver
- Intel E5-2698 v3 @ 2.30GHz

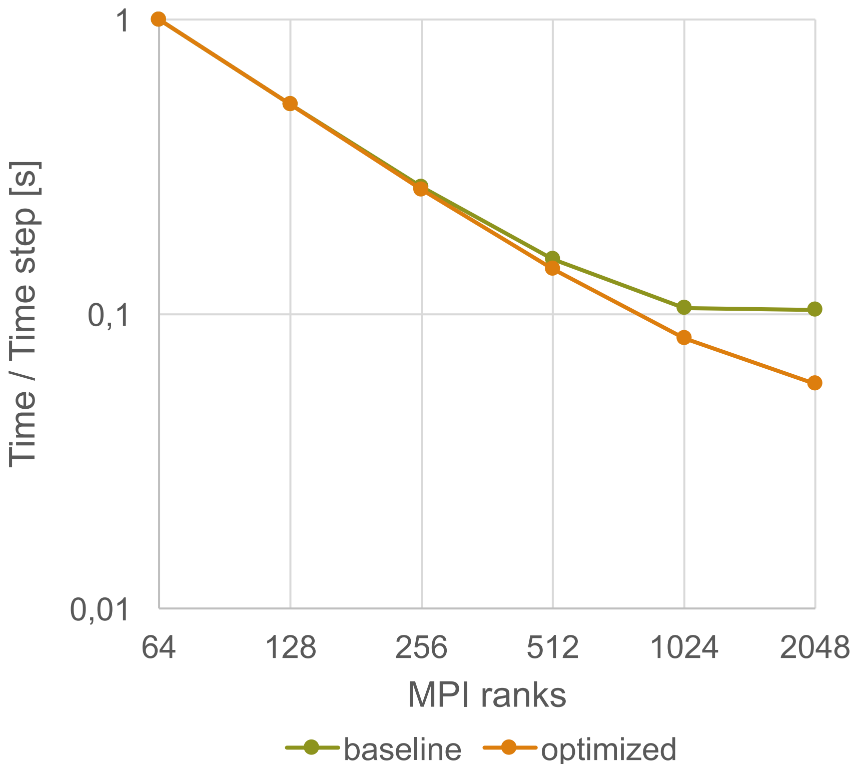
- 3 weeks of work…



Scalability Results

- OpenFOAM/2.2.2
- ~100M cells
- GAMG solver
- Intel E5-2698 v3 @ 2.30GHz

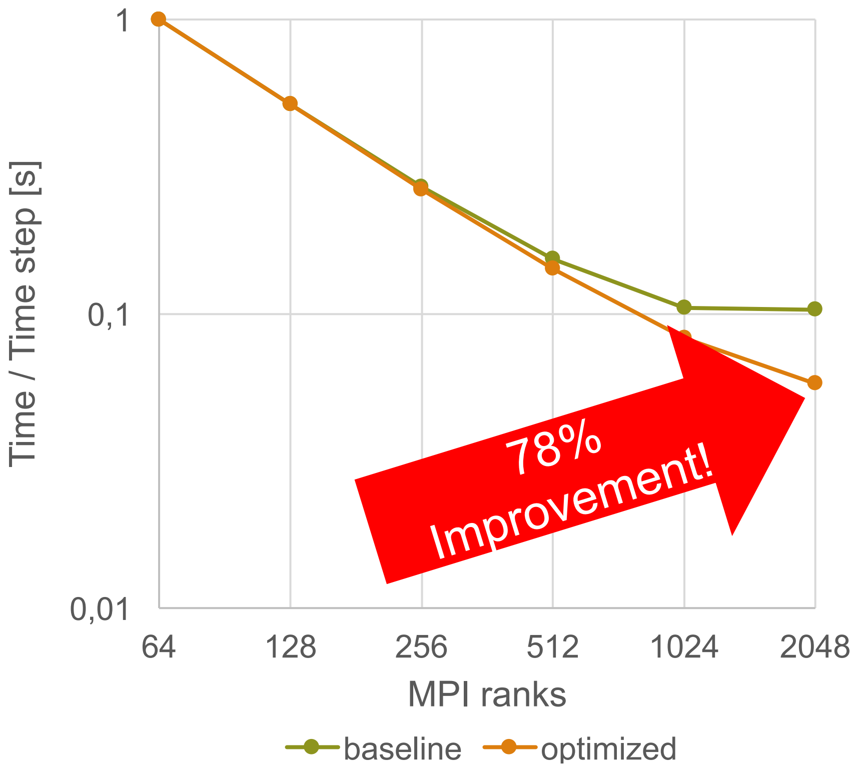
- 3 weeks of work...



Scalability Results

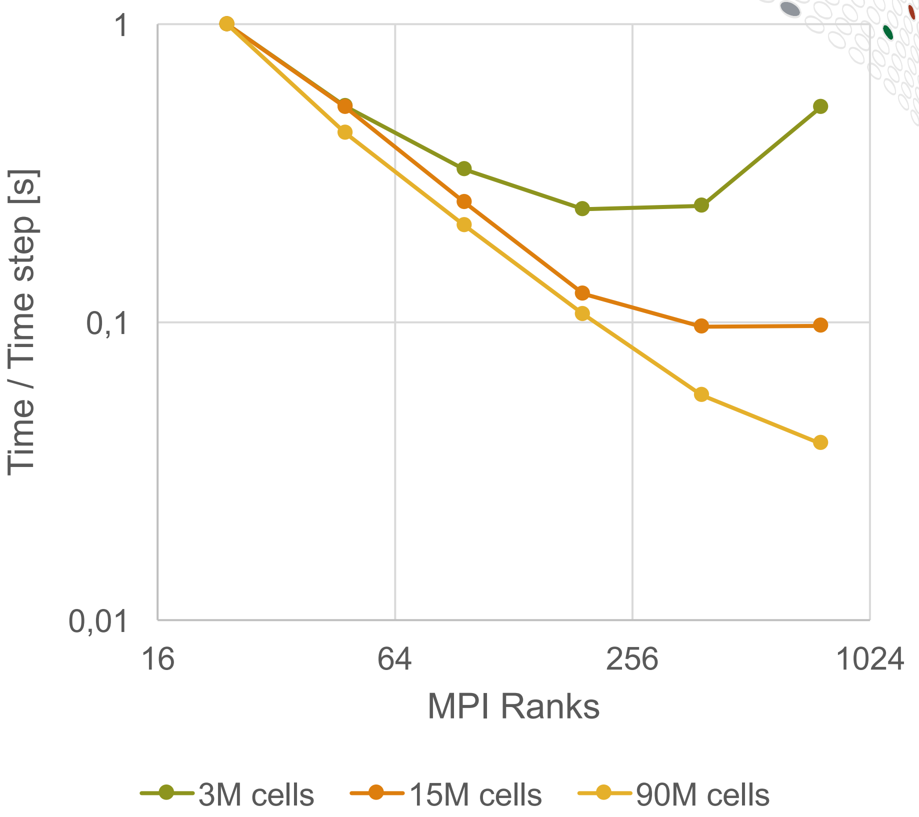
- OpenFOAM/2.2.2
- ~100M cells
- GAMG solver
- Intel E5-2698 v3 @ 2.30GHz

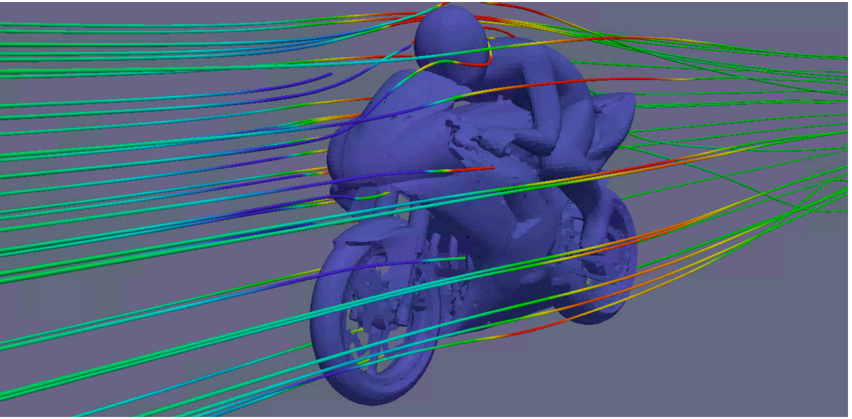
- 3 weeks of work...



Scalability Limit

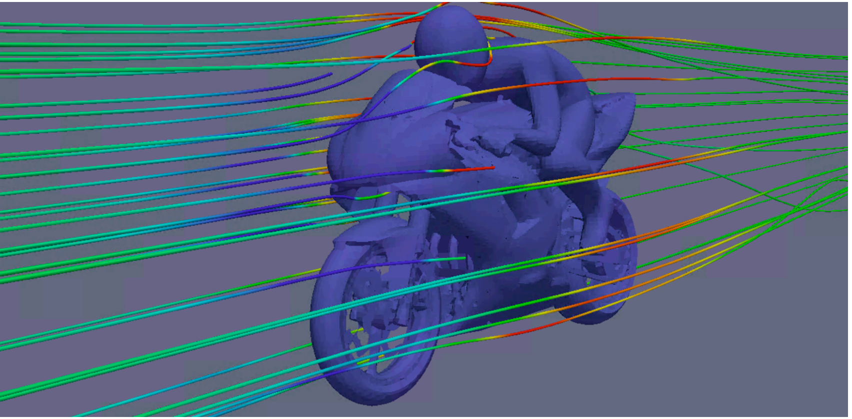
- Inflated motorBike
- Intel E5-2680v3 @ 2.5GHz
- OpenFOAM/2.3.1

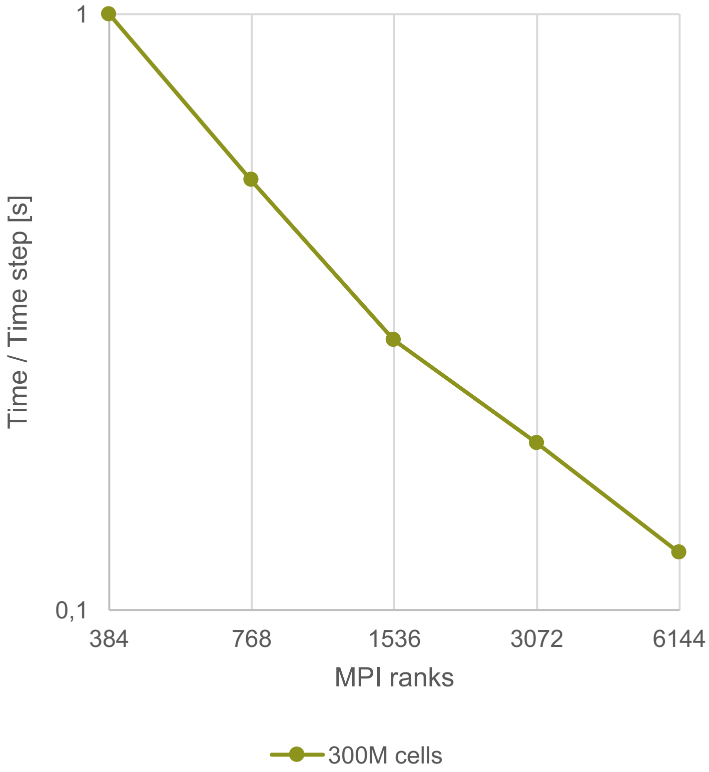




Scalability Limit

- Inflated motorBike
- Intel E5-2680v3 @ 2.5GHz
- OpenFOAM/v1612+



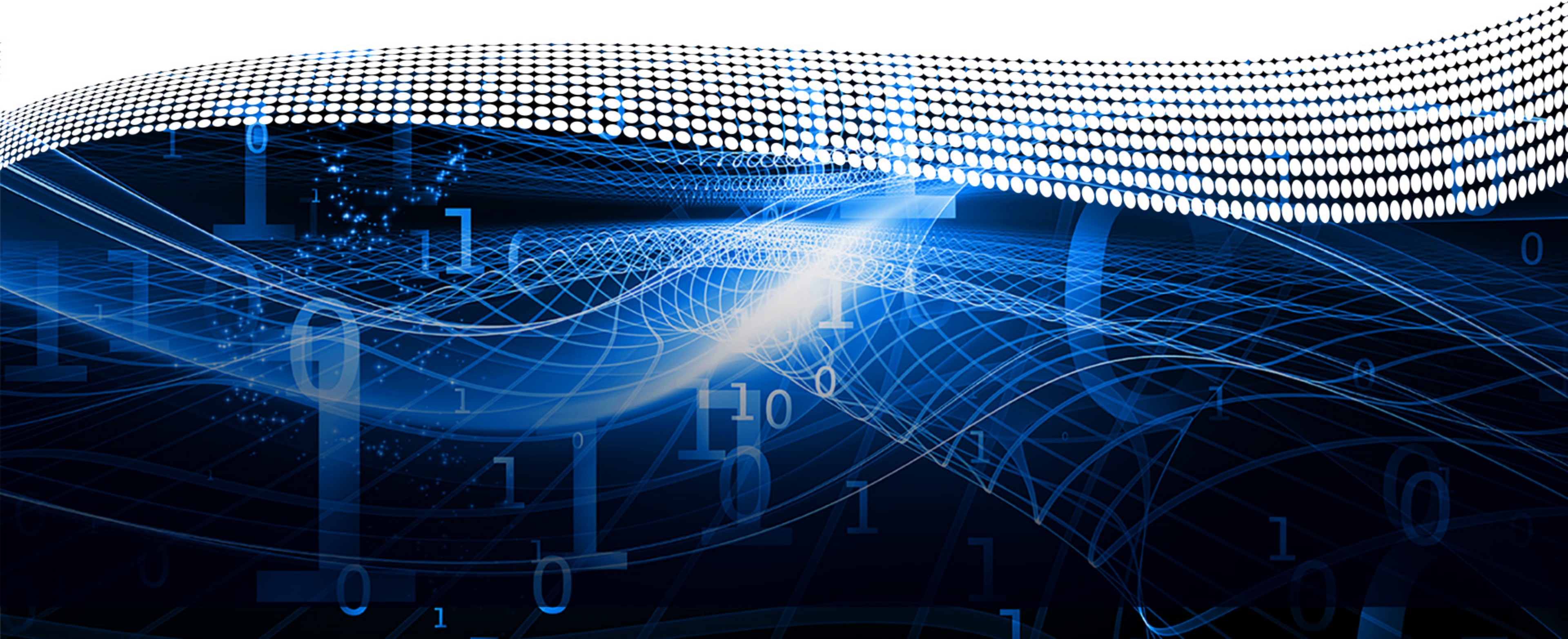


Recap | Where do we want to go?

- **Multi and many core architectures**
  - A lot more cores to feed
  - Need for further scaling and/or hybrid approach
- **Wider SIMD/Vector instructions**
  - Suboptimal vectorization will hurt you more
- **Find optimal solution for I/O design**
  - This may be solved from vendor side

CRAY®

COMPUTE | STORE | ANALYZE



Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.