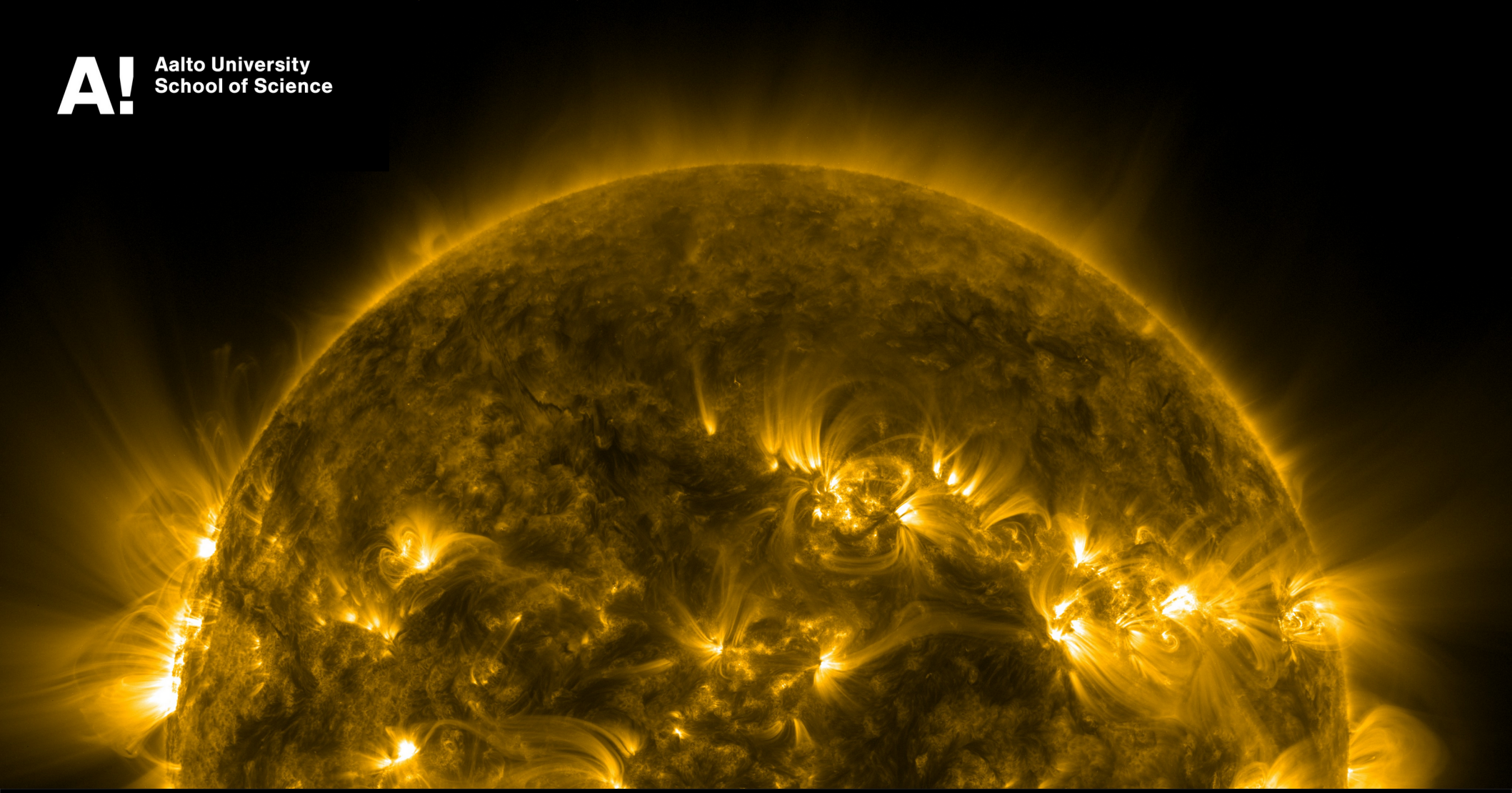
Aalto University
School of Science



Astaroth and the future of HPC

Pencil Code User Meeting, 2021, May 21st

Johannes Pekkila

Image credit: NASA/SDO

Speaker

- Johannes Pekkilä
  - Doctoral Student at Aalto University, Finland
  - Field: Computer Science (Big Data and Large-Scale Computing)
  - Worked on accelerating physical simulations on GPUs since 2014
  - Thesis subject: accelerating computational patterns in HPC

Astaroth

- A multi-node GPU library for stencil computations
- Controlled via a C API
  (Python & Fortran bindings available)
- A domain-specific language for adding physics
  (Performance & productivity)
- Tuned for computational physics
  (Very high cache-efficiency in comparison to competitors)

This talk

1) Motivation and the current state of Astaroth

2) Future of Astaroth

3) Future of HPC

Why GPUs?

- A GPU has ~10x higher throughput than a CPU
  (FLOPs, memory bandwidth)
- An HPC node typically houses more GPUs than CPUs
  (CSC's Puhti: 4x GPUs and 2x CPUs per node)
- Hybrid CPU-GPU machines beat CPU-only machines in performance
  (CSC's LUMI ~550 PFLOPS, Mahti ~8 PFLOPS)

Single-node performance

Full MHD, double precision

Pencil Code on 2x Intel Xeon Gold 6230 CPUs, 40 cores:
- 23 ns / grid point / step

35x effective speedup

Astaroth on 4x Tesla V100 GPUs
- 0.65 ns / grid point / step

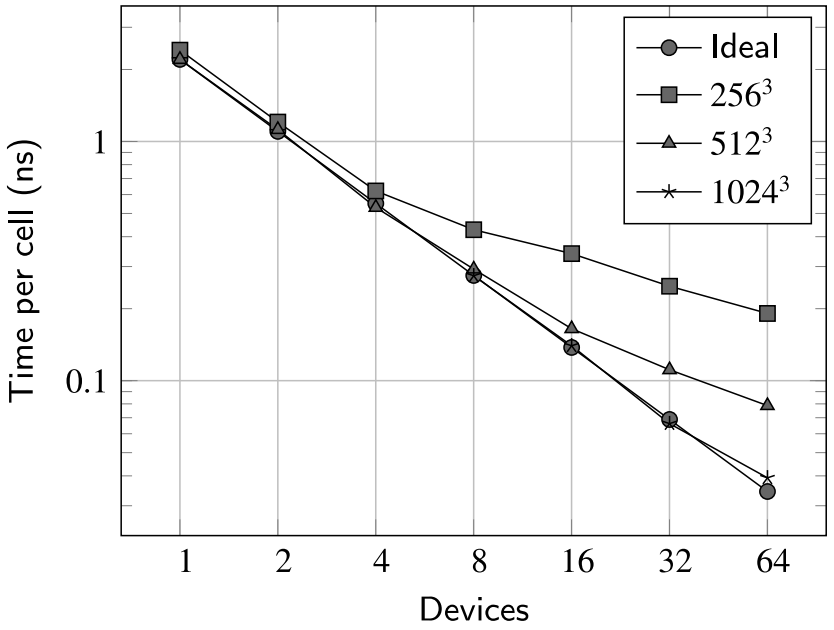
Details: Vaisala 2021, https://iopscience.iop.org/article/10.3847/1538-4357/abceca

Disclaimer: the theoretical per-node speedup is 16x (CPU vs. GPU memory bandwidth). We suspect that PC leaves some performance on the table.

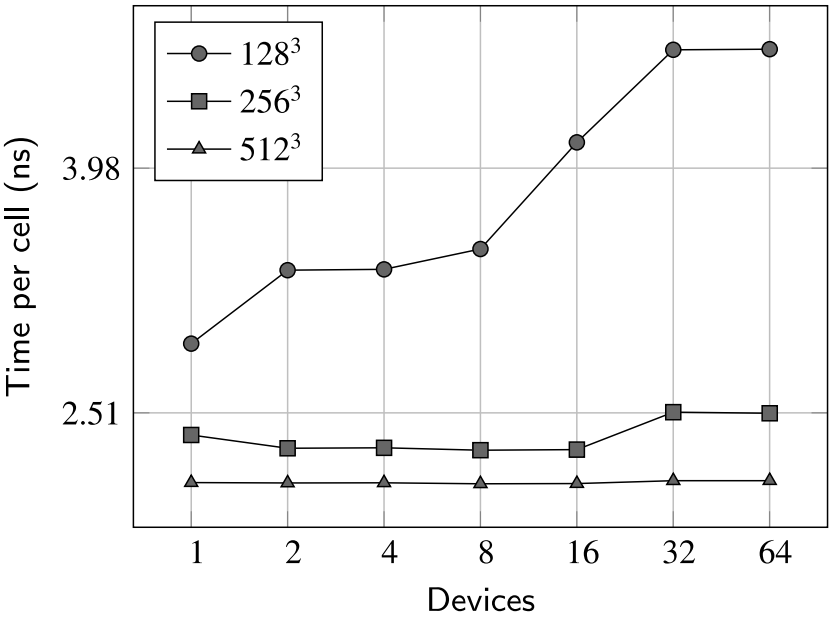
Multi-node performance

Full MHD, double precision

Strong scaling



Weak scaling



For details on the basic MPI implementation, see https://arxiv.org/abs/2103.01597
Scaling efficiency further improved with a task scheduler by O. Lappi in 2021, see
http://urn.fi/URN:NBN:fi-fe2021050428868

FAQ

- Can’t we just run on more CPUs?
  Yes, but you’ll use 10x more billing units and wait longer in queue.
  Additionally, hybrid GPU systems beat CPU-only systems in full-machine runs.

- Do GPUs support double precision? Is it efficient?
  Yes and yes.

FAQ cont'd

- Don't GPUs have too small memories compared to CPUs?
  Yes, but GPUs have been catching up. For example, a Puhti node has 384 GiB CPU memory and 128 GiB GPU memory.

- CPU-GPU communication via the PCIe bus is slow and will kill the performance
  This is mostly true, but applies only when accessing small data segments randomly. Sufficiently large, predictable data transfers can be carried out efficiently by pipelining CPU-GPU transfers and using GPUDirect RDMA.

This talk

1) Motivation and the current state of Astaroth

2) Future of Astaroth

3) Future of HPC

Astaroth on AMD

LUMI pilot on December 2021

- CSC's LUMI stated to "*provide ~550 PFLOPS, mostly from AMD GPUs*". Exact details not yet public.
- Astaroth is one of the pilot projects
- Very high-res runs planned, up to 16k cubed

Astaroth on AMD

LUMI pilot on December 2021

- AMD support added to Astaroth in the latest development branch
- AMD-specific optimizations currently work in progress

Further improvements

- ☑ AMD support
- ☐ AMD optimization
- ☐ DSL improvements
- ☐ Test-field methods, spherical coordinates, etc
- ☐ Further R&D (compression, deep learning, FPGAs)

This talk

1) Motivation and the current state of Astaroth

2) Future of Astaroth

3) Future of HPC

Future of HPC

- GPUs for scientific computing have started to diverge from pure graphics accelerators to data-processing units (DPUs)
- Larger caches, better f64 performance, more parallelism
- NVIDIA's ARM-based HPC CPUs (not a typo) to alleviate CPU-GPU interconnect bottlenecks
- Silicon-based microprocessors expected to hit a performance wall 2025 onwards: expect disrupting technologies to emerge
- Speculative: FPGAs could offer a competitor to GPUs. Microsoft's Project Catapult, Brainwave, AMD's acquisition of Xilinx, Intel's acquisition of Altera

Big Picture

- Compute is cheap, bandwidth is expensive
- The gap between arithmetic performance and memory bandwidth continues to widen on future hardware
- Need to do everything we can to reduce data transfers
- Need to ask hard questions: are we ready to trade off accuracy for speed? Lossy compression
  (see f.ex. Lidstrom, 2014, https://doi.org/10.1109/TVCG.2014.2346458)
- Interesting state-of-the-art ideas: physically guided reconstruction of noisy data with deep learning
  (see f.ex. Kim et al., 2021, https://doi.org/10.1017/jfm.2020.1028)

Conclusion

- Astaroth is a highly efficient and scalable multi-node library for stencil computations on NVIDIA and AMD GPUs. Tuned for requirements in computational physics
- Built-in support for full MHD, RK3, and 2^{nd}, 4^{th}, 6^{th}, 8^{th} order finite differences
- We expect to run very high-resolution runs on LUMI this year
- Expect disruptions in HPC hardware in the next 10 years
- Open question: what can we do about data transfer bottlenecks?

Astaroth

Code
http://bitbucket.org/jpekkila/astaroth

Publications

DSL (Pekkila 2019) http://urn.fi/URN:NBN:fi:aalto-201906233993

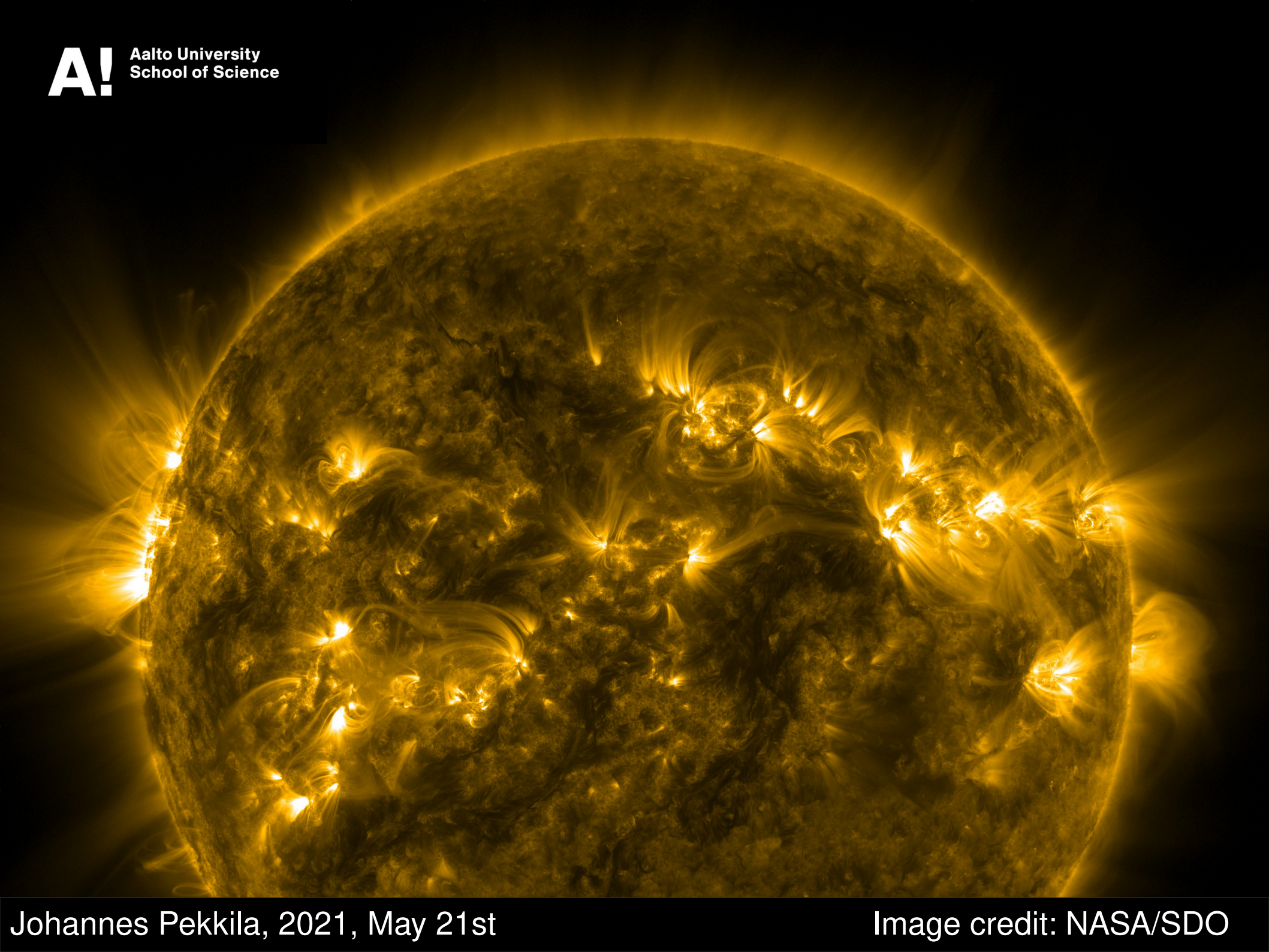
Physics (Vaisala 2020) https://arxiv.org/abs/2012.08758

MPI (Pekkila 2021) https://arxiv.org/abs/2103.01597

Task Scheduler (Lappi 2021) http://urn.fi/URN:NBN:fi-fe2021050428868

Astaroth

Special thanks
M. Väisälä
M. Käpylä
M. Rheinhardt
O. Lappi
Aalto University Astroinformatics Group
CSC – IT Center for Science
ERC Consolidator Grant UniSDyn


Aalto University
School of Science
Johannes Pekkila, 2021, May 21st
Image credit: NASA/SDO