CATALYST
Leveraging HPC to Drive Innovation in AI

HLRS' Strategy towards a Convergence of HPC and AI

Dennis Hoppe (HLRS)

AI Strategy@HLRS

CHALLENGES OF AI

European Strategy for AI — Three Pillars [1]

- Boosting the EU's technological and industrial capacity and **AI uptake across the economy**
  - Supporting AI research excellence centres across Europe
  - Bringing AI to all small businesses and potential users
- Preparing for **socioeconomic changes**
  - Focus on jobs that are likely to be transformed or to disappear; leverage chances of new job creations
- Ensuring an appropriate **ethical and legal framework**
  - Citizens and businesses alike need to be able to trust the technology they interact with

[1] European Commission: Communication Artificial Intelligence for Europe, 2018.

Current AI Initiatives in Europe

- European AI Alliance [2]
  - Involve all stakeholders within Europe that are affected by AI
  - Dedicated platform where they can offer input and feedback to the high-level expert group on AI
- High-Level Expert Group on Artificial Intelligence [3]
  - Works on ethic guidelines towards "**Trustworthy AI**"
  - Steering group of the European AI Alliance
- **AI On-Demand Platform** [4]
  - Comprehensive European AI-on-demand platform to
    - lower barriers of innovation
    - boost technology transfer
    - catalyse the growth of start-ups and SMEs

[2] DG Connect: The European AI Alliance, 2019.
[3] Robotics and AI Group of the EC: High-Level Expert Group on Artificial Intelligence, 2019.
[4] Thales SAS: AI On-Demand Platform (ai4eu.eu), 2019.

Why does AI need HPC? Why does HPC need AI?

- **AI solutions require immense compute-resources**
  - CPU, network, storage, accelerators, …
- **Simulations** such as climate models are **hitting the wall**
  - Computing physical processes right down to the last detail is very compute-intensive
- **Information overload** will continue to increase
  - 5G, IoT, autonomous driving and flying, …
- HLRS addresses these challenges through different channels
  - **Economy**, **Society**, **Research**

AI@HLRS

- **Economy** (with focus on SMEs)
  - Lacking AI expertise
  - No in-house AI hardware
  - Security concerns / data mgmt. (GDPR)
- **Society**
  - AI is seen as a blackbox model
  - Low acceptance rates of AI solutions
  - Security concerns (privacy intrusion)
- **Research**
  - Support of hybrid HPC/AI workflows on HPC systems
  - Resolve multitude of complementary requirements (e.g. software)
  - Interdisciplinarity: AI experts are no HPC experts

Combining HPC and HPDA for Academia and Industry

CATALYST

The Catalyst Project [2016–2021]

- Our customers tend to run more and more **data-intensive applications** resulting in **vast amounts of output data**
  – A single turbulence & acoustics simulation of an axial fan with just four rotations results in 80 TB of data
  – Domain experts are no longer able to analyse data manually
- Close cooperation between **HLRS** and Cray (→ **HPE**)
- Evaluate requirements that arise when combining AI and HPC
  – Hardware + software environment
    - Cray Urika-GX (DA/ML), CS-Storm (DL), HPE Apollo (HPC)
    - Build upon open-source software stack
  – Perform **case studies** with both academia and industry

https://www.hlrs.de/bigdata

Case Studies

- **Speech2Text models for German language** (LandesCloud)
- **Formal verification of neural networks** (Fraunhofer IPA)
- **Deep reinforcement learning for robotics** (Festo)
- **Material characterization for metal forming** (University Stuttgart)
- **Smart alerting system for freezers** (CHECK)
- **Identification of trends in scientific publications** (Leichtbau BW)
- **“3D City over Night“** (nFrames)
- **Data analytics summer school** (HS Alb-Sig)
- **Prediction of S-Bahn delays** in Stuttgart (HLRS)
- **SmartSHARK** (University of Goettingen)
- **Performance variations in HPC jobs** (HLRS)
- **Turbulence detection in air flows** (RWTH Aachen)
- **Complications with biomechanical devices** (HLRS)
- …

Scenario A: Processing of Massive Datasets

- With the improvements in system performance, HPC users are able to run
  - more simulations in the same time,
  - more complex simulations (e.g. finer meshes),
  - and thus **produce massive amounts of data!**
- Data produced can no longer be manually analyzed
  - requires domain experts and manual inspection

➢ **Let AI and DA automatically analyse data** to reveal interesting insights hidden within the data!

Case Study: “3D City over Night” (nFrames)

The illustration shows a textured 3D mesh of San Francisco. The data was provided by courtesy of Geomni. Copyright nFrames.

O. Shcherbakov et al. URIKA-GX PLATFORM'S MULTI-TENANCY:LESSONS LEARNED, CUG 2019.

Scenario B: Parameter Sweeping via Feedback Loops

- **Let AI models optimize the parameter space** between simulations to reduce the overall number of required application runs [6]
  - Use output data from previous simulations to predict "better" input parameters to be used in future simulations
  - **Drop simulations that are likely to yield similar results**
- Advantages for the
  - User
    - saves time and resource costs
  - HPC center
    - saves energy
    - frees up resources to be allocated to other users

[6] Silva et al.: JobPruner: A Machine Learning Assistant for Exploring Parameter Spaces in HPC Applications, CoRR '18.

Scenario C: AI at the Edge

- Leverage full potential of **Internet of Things** via **edge computing**
  - Edge devices are becoming more powerful
  - Collect, preprocess, and analysis of (streaming) data
  - Avoid delays from sending everything into the Cloud
- Eliminate most communication with Cloud/HPC by making edge devices smarter: AI@edge
  - E.g. detect anomalies, generate predictions on the fly
  - A perfect fit for tasks such as predictive maintenance
- Typical workflow
  - 1) Run compute-intensive training of AI models on HPC
  - 2) Perform light-weight inferences at the edge

Scenario D: Mixed Workloads

- **Let AI speed up simulations**
  - AI-based models are able to replace computationally intensive-tasks in simulations
    - e.g. compute-intensive Monte Carlo simulations can be exchanged with a more light-weight trained AI model [7]
- **Let AI improve simulation accuracy**
  - Exploit AI methods to model physical aspects that are currently too complex to be understood entirely
    - e.g. effects of cloud formation are such a complex problem
    - meshes used for simulations are too coarse to model clouds
    - automatic grid optimisation during simulations through AI [8]

[7] Baydin et al.: Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model, CoRR '18.
[8] Rasp et al.: Deep learning to represent subgrid processes in climate models, PNAS '18.

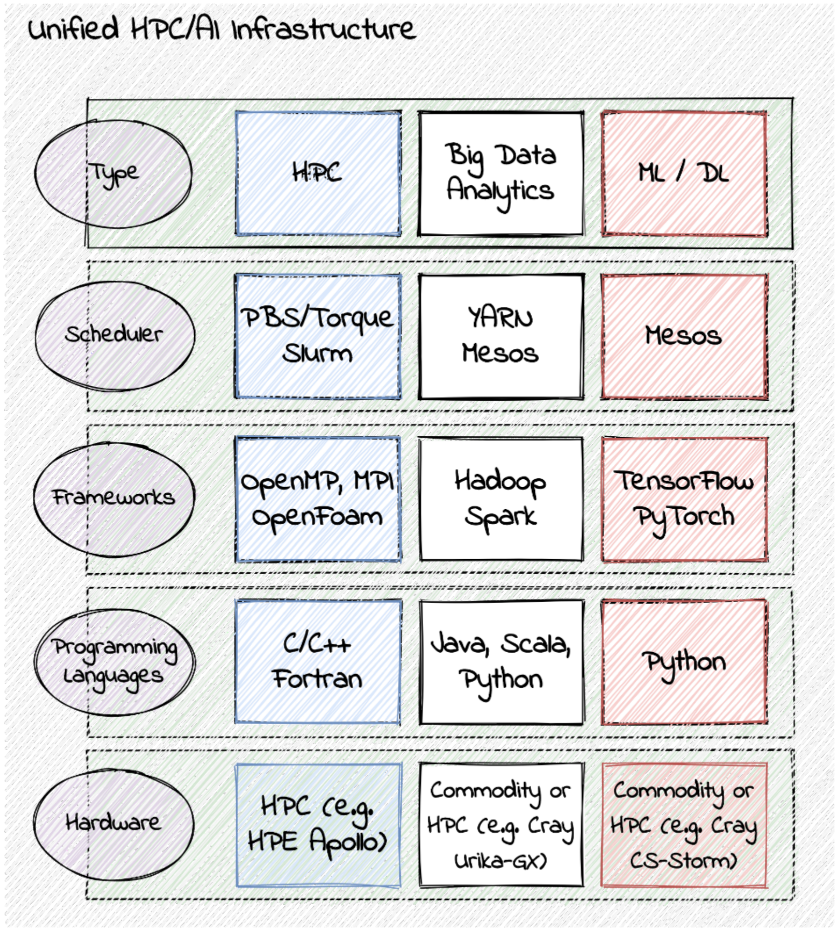
Why does Artifical Intelligence require HPC?

How can simulations benefit from AI?

CONVERGENCE OF HPC AND AI

Technical Challenges

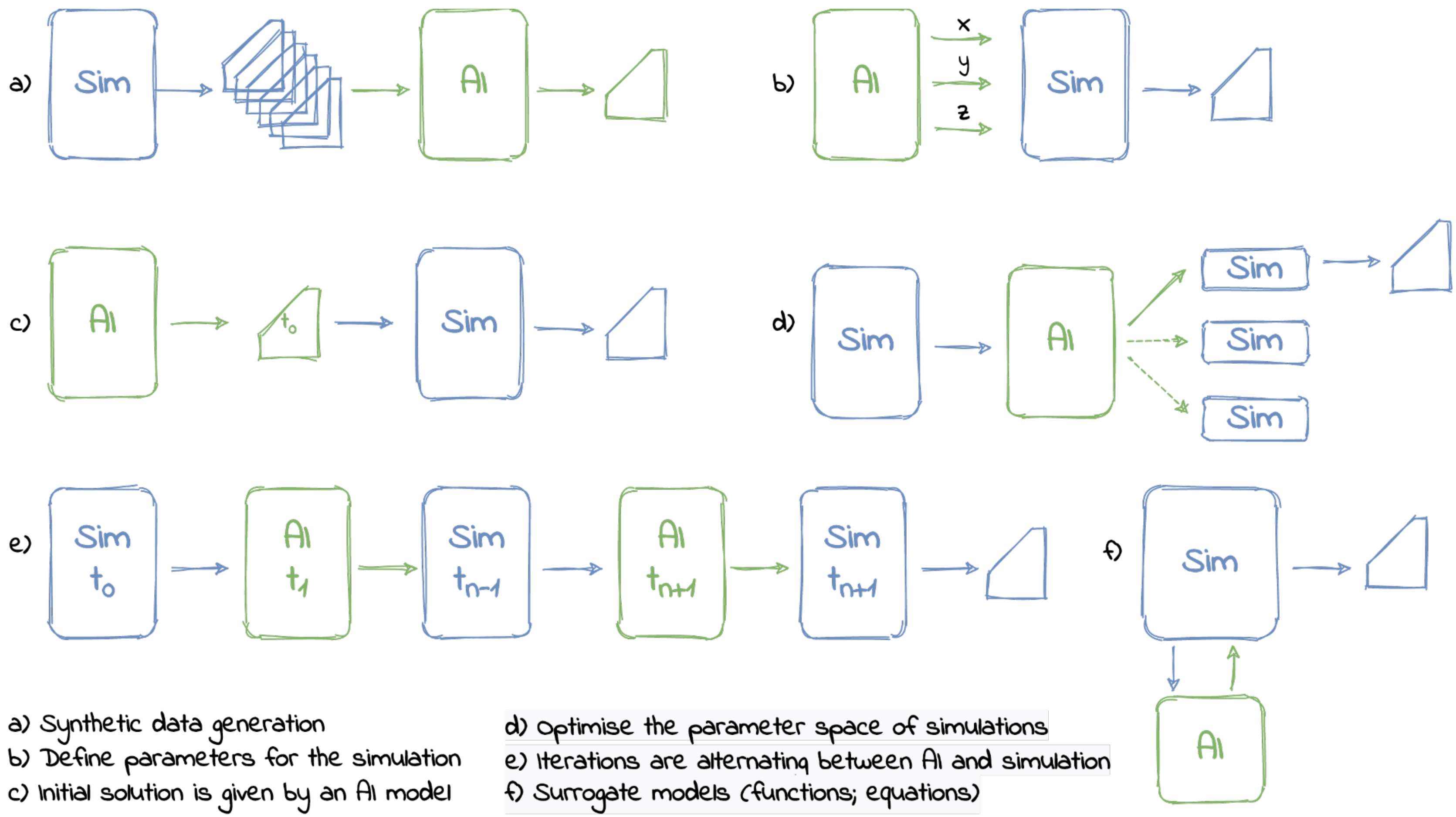
- User needs
  - Combine and/or integrate AI into typical HPC workflows
- We should have an **unified software stack** and, ideally, a **single resource manager** to deploy AI workloads onto HPC
- Challenges and drawbacks
  - cf. next slide



Technical Challenges (cont'd)

- Allow **DA/ML/DL** frameworks to run on HPC systems
  - **Containerization** is the way to go (e.g. via Singularity)
  - Provide an **holistic resource manager** to run HPC, data analytics, machine learning and deep learning jobs
    - integration with PBS Pro, for example
  - Introduce **streaming processing** to HPC (IoT)
- **Large-scale AI**
  - Improve acceleration and scalability of AI on HPC
    - e.g. via offloading through RDMA (e.g. TensorFlow, Spark)
- Work on specific examples coming from the engineering domain to **showcase benefits** of the convergence of AI and HPC

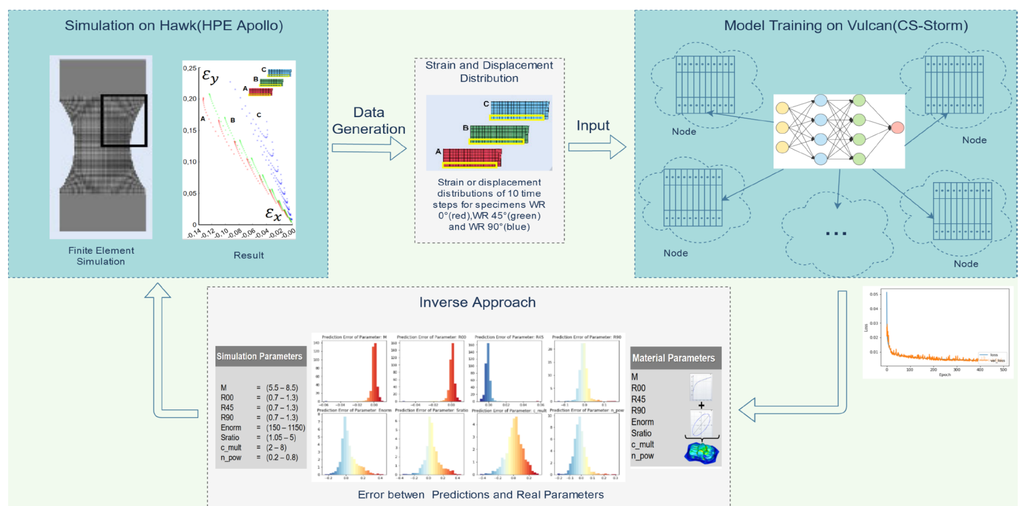
Examples of Hybrid HPC/AI Workflows



a) Synthetic data generation
b) Define parameters for the simulation
c) Initial solution is given by an AI model
d) Optimise the parameter space of simulations
e) Iterations are alternating between AI and simulation
f) Surrogate models (functions; equations)

Case Study: Distributed Learning on GPUs

- Problem: Material Characterization of Sheet Metals
  - Sheet metal forming processes require material parameters as input
  - Validation is very time-consuming (inverse parameter identification)
- Solution: Combination of FEM and DNN
  - Replace the time and compute-intensive inverse approach by DNN model to perform material parameter validation much more efficiently
    - **Phase 1:** FEM simulation generates synthetic data
    - **Phase 2:** Train a DNN model on the data to predict material parameters



Take Away Message

- **AI Strategy of HLRS** aligns well with the European one
  - Address societal challenges (e.g. ethics)
  - Support SMEs to work together on research problems
  - Push the convergence of AI and HPC; hybrid workflows
- The future of HPC requires a system architecture to run HPC, data analytics, machine and deep learning workflows **on the same system as part of a complex workflow** [9]
- Advancements of the AI software stack is required to leverage the full potential of HPC
  - Incorporation of container technologies into HPC (e.g. singularity)
  - Scaling of frameworks such as Spark (e.g. via RMDA support)
  - Interplay with shared file systems (e.g. Lustre) since AI frameworks are optimized for data locality

[9] IDG Communications: 7 Drivers for HPC and AI Convergence, 2019.

Thank you !

Questions ?

Dennis Hoppe
High Performance Computing Center Stuttgart
Nobelstraße 19
70569 Stuttgart

Federal Ministry
of Education
and Research

This work was supported by the research project CATALYST funded by the Ministry of Science, Research and the Arts of Baden-Württemberg, Germany (2016–2021).