Smart Networks in HPC
An Architecture Perspective

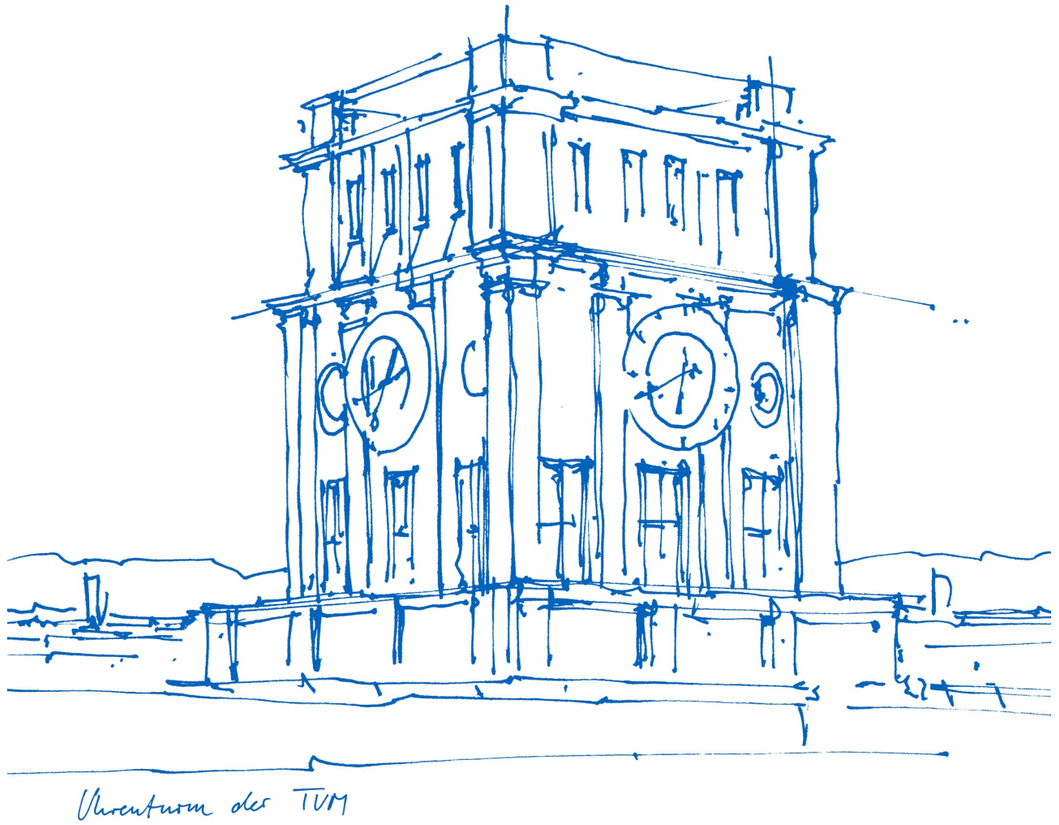
Challenges and Opportunities

Martin Schulz

Chair for Computer Architecture and Parallel Systems
Technical University of Munich (TUM)

International Workshop on Smart Networks,
Data Processing and Infrastructure Units

Thursday June 25th, 2023



Smart Networks in HPC
*An **Inverted** Architecture Perspective*

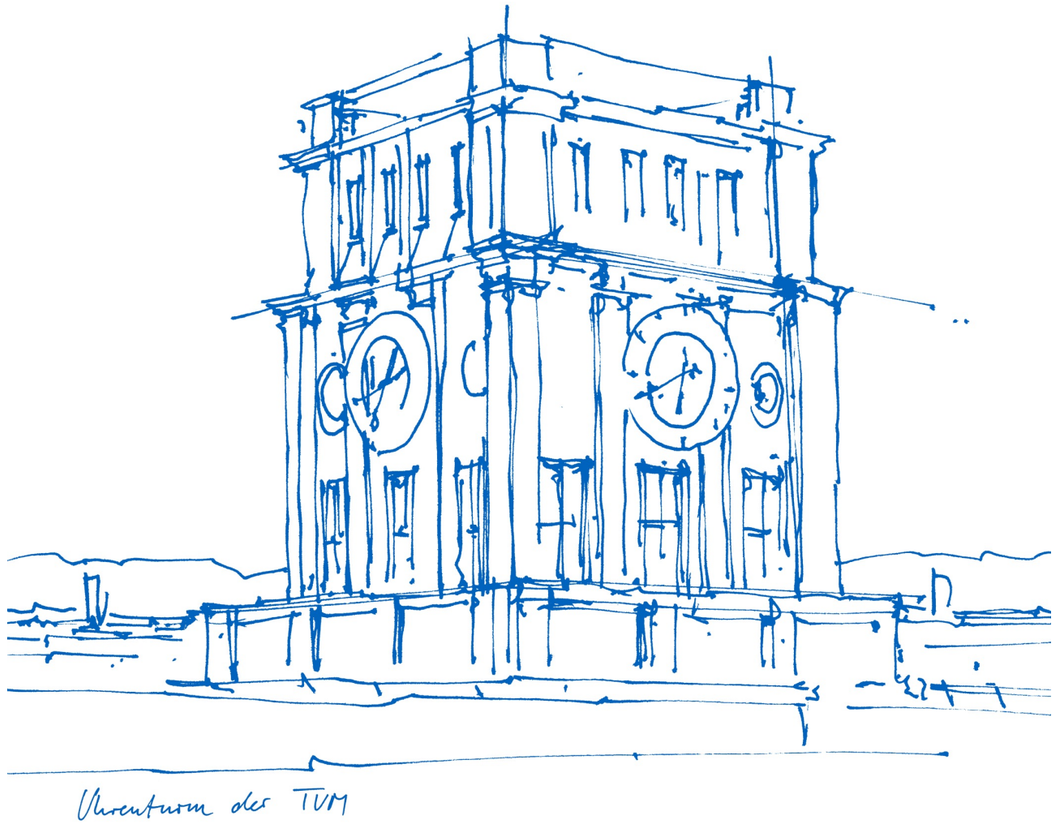
Challenges and Opportunities

Martin Schulz

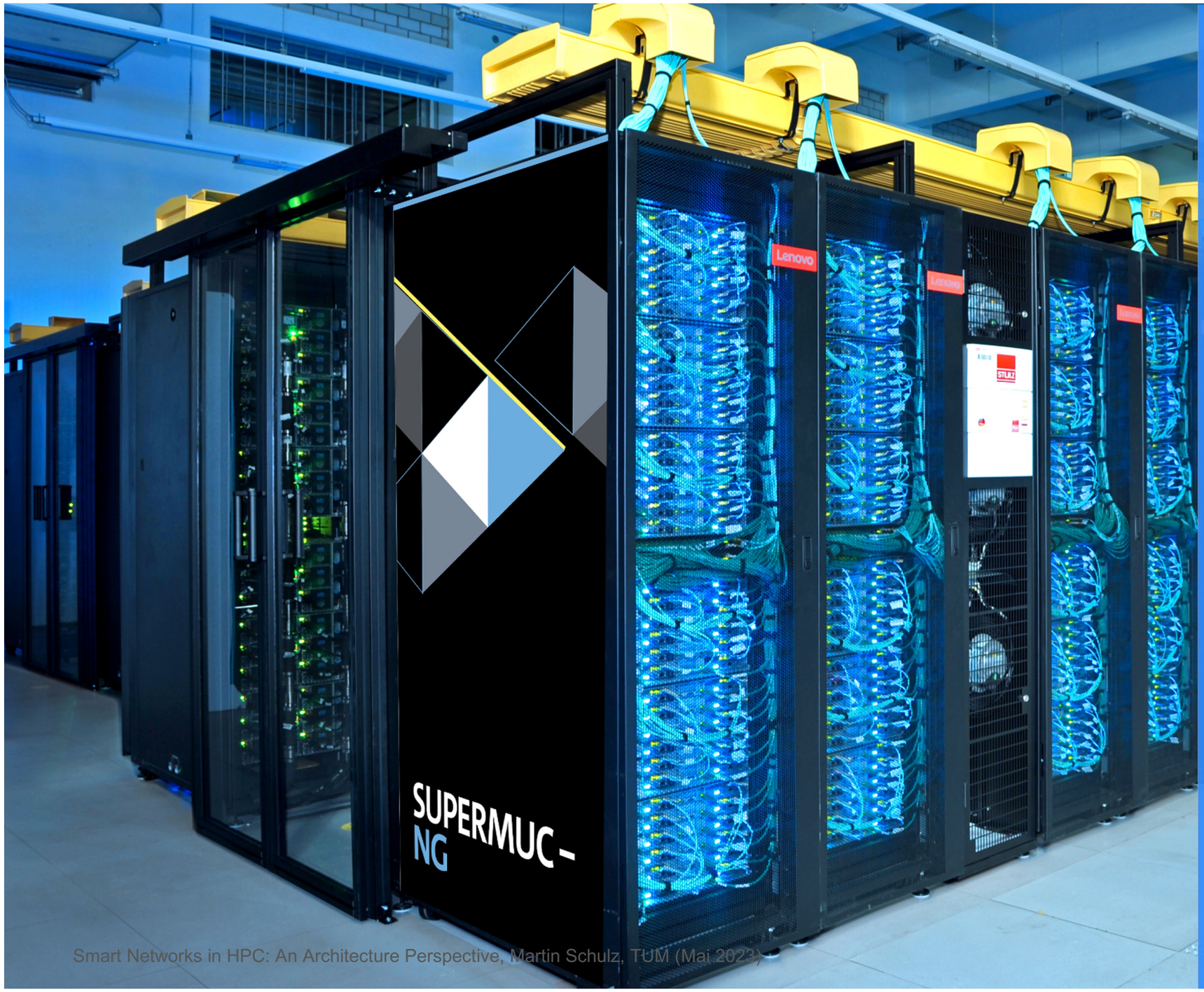
Chair for Computer Architecture and Parallel Systems
Technical University of Munich (TUM)

International Workshop on Smart Networks,
Data Processing and Infrastructure Units

Thursday June 25th, 2023




Department of Computer Engineering
Technical University of Munich
Leibniz Supercomputing Centre (LRZ)
of the Bavarian Academy of Sciences



SuperMUC-NG

Top500 (June 2023): #31

Lenovo Intel (2018)

311,040 cores
Intel Xeon Skylake

26.9 PetaFlops
Peak

719 TeraByte
Main Memory

70 PetaByte
Disk

Broad application mix, deep in-house expertise in supporting scientific discovery

A First in European Procurement

The ExaMUC Project

SuperMUC-NG Phase 2 currently rolling in

- Lenovo/Intel System
- Accelerated nodes using Intel's Ponte Vecchio
- **More performance with less power on significantly less footprint**

Already planning the next: ExaMUC

- Launched in Jan. 2022 and installation in ~2025
- Co-Design process

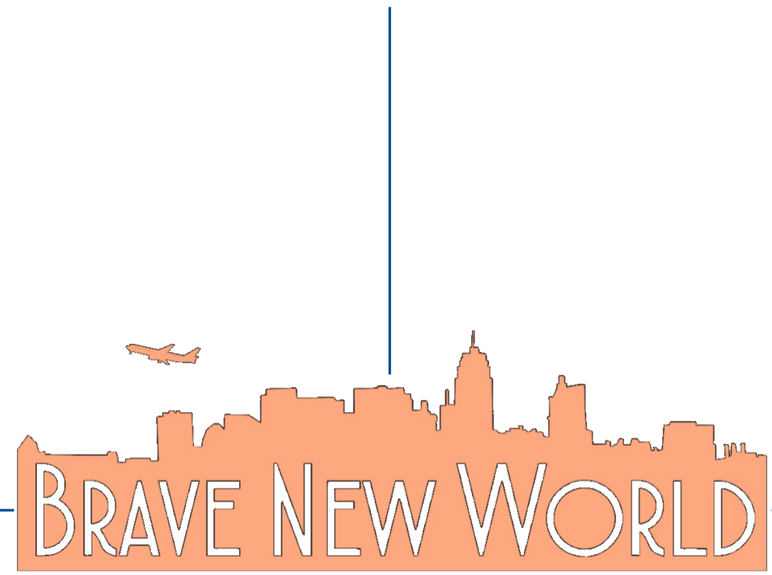
Innovation Partnership

- Phase 1: prototype development (hardware and software) with multiple vendors
- Gain innovations for both sides and enable common learning
- Phase 2: selection of actual ExaMUC machine

Trends in HPC

Power Limits

Specialization



Data Movement = Cost

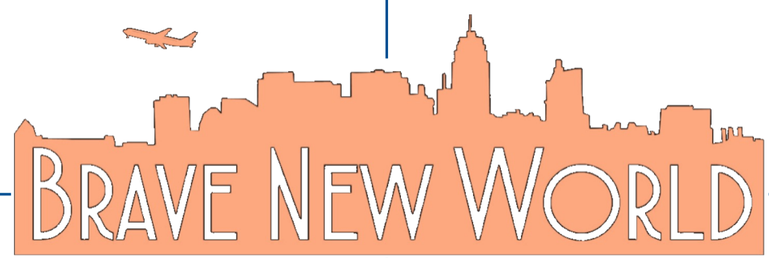
More Diverse Workloads

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Trends in HPC (with a Network View)

Power Limits

Specialization



Data Movement = Cost

More Diverse Workloads

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Power and Energy as Hard Constraints

Power Limits

Cost are limiting

Power
pushing

Societal pressure

Specialization

Data Movement = Cost

More Diverse Workloads

BRAVE NEW WORLD

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Cambrian Explosion of Architectures

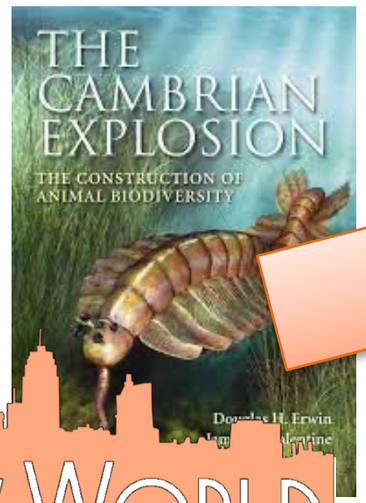
Power Limits

Cost are limiting

Power (
pushing

Societal pressure

Very Little Power Control



Hidden

Specialization

Driven
g and
g end of Moore's Law

Feature reduction is ending

Data Movement = Cost

More Diverse Workloads

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Cost of Data Movement is Becoming a Limiter

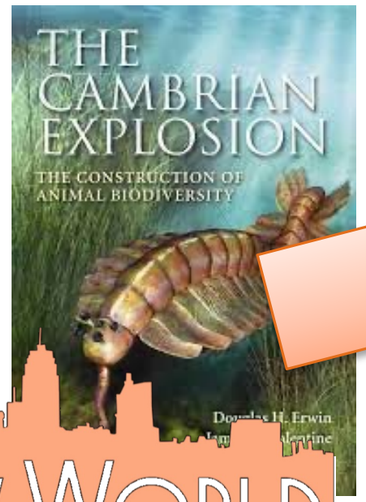
Power Limits

Cost are limiting

Power (
pushing

Societal pressure

Very Little Power Control



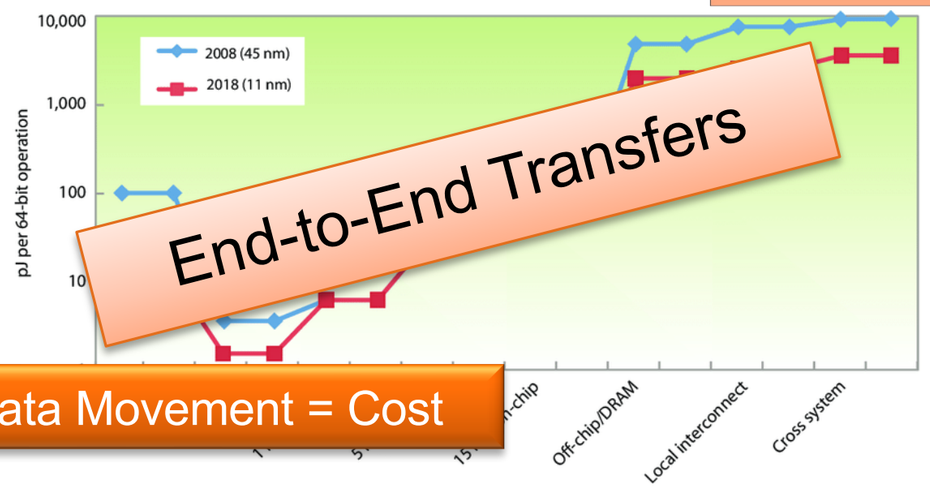
Hidden

Specialization

Driven
g and
g end of Moore's Law

Feature reduction is ending

BRAVE NEW WORLD



Data Movement = Cost

More Diverse Workloads

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Workloads are Becoming More Complex and Diverse

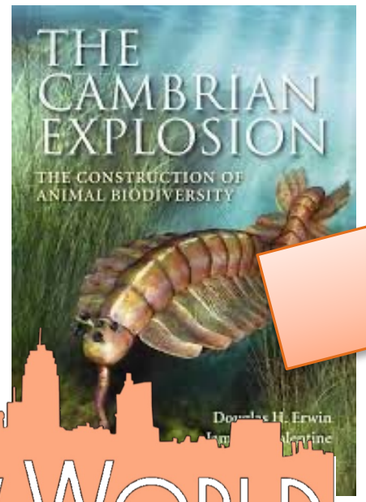
Power Limits

Cost are limiting

Power (
pushing

Societal pressure

Very Little Power Control



Specialization

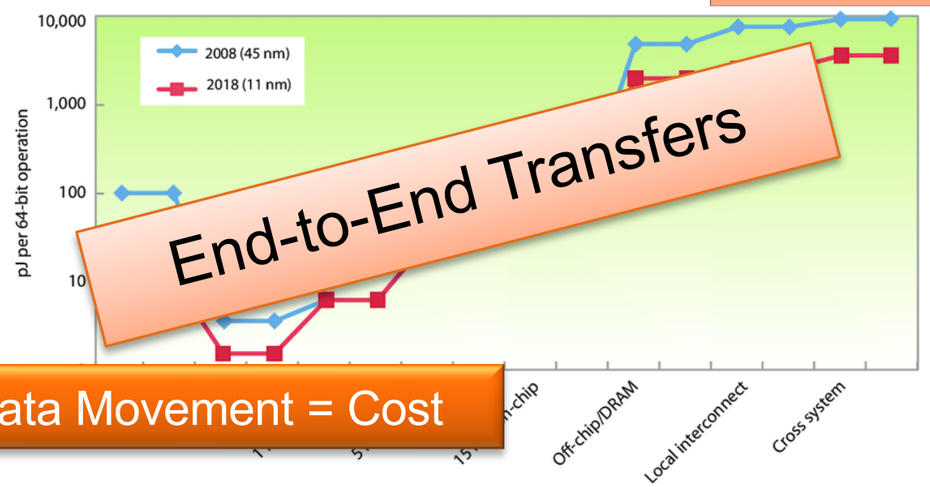
Driven

Hidden

g end of Moore's Law

Feature reduction is ending

BRAVE NEW WORLD



End-to-End Transfers

Data Movement = Cost

One Size Fits All

More Diverse Workloads

Smart Networks in HPC: An Architecture Perspective, Martin Schulz, TUM (Mai 2023)

Networks Need to Change Going Foward

Smart networks offer opportunities

- Processors or reconfigurable logic on the NIC
- Processing power in switches

NVIDIA
Bluefield

Ability to offload simple tasks

- MPI protocol processing
- Monitoring, correctness checking

Silcom N5010 Series

Advantages

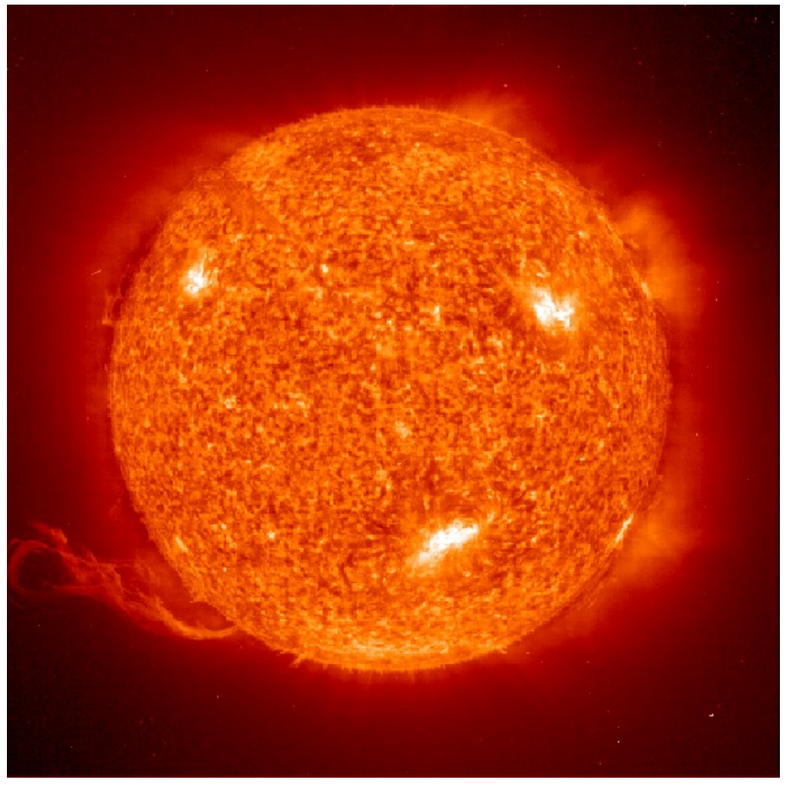
- Possible energy reduction
- Spezialization of the NIC
- Reduction in data transferred possible
- Customization to workloads

APS Networks
Programmable Switches

Optimizing Plasma Codes in DarExa-F

(Partners: MPG MPCDF, MPG IPP, TUM, FAU, ParTec, LRZ)

SPONSORED BY THE
Federal Ministry of Education and Research



Codes for Plasma Physics

- Target: processes as in ITER
- Gene Family of codes

Gene is highly optimized scalable, but

- Currently still fixed on double precision
- Subject to large data transfers

DPUs can help overcome this issue

- Compression on the fly
- Combined with reductions
- Reduced data volume for faster and more enery efficient communication
- Transparent to application (as part of MPI)

Challenges

Extensions of programming interface

- Integrate coordinated computation
- Manage multiple binaries transparently
- Ideally transparent to the application

Approach: MPI Extensions

- Drive certain communication from NIC
- Coordinate different data paths
- Controlled via hints on communicators
- Ability to enable via PMPI interface

Most important challenge:

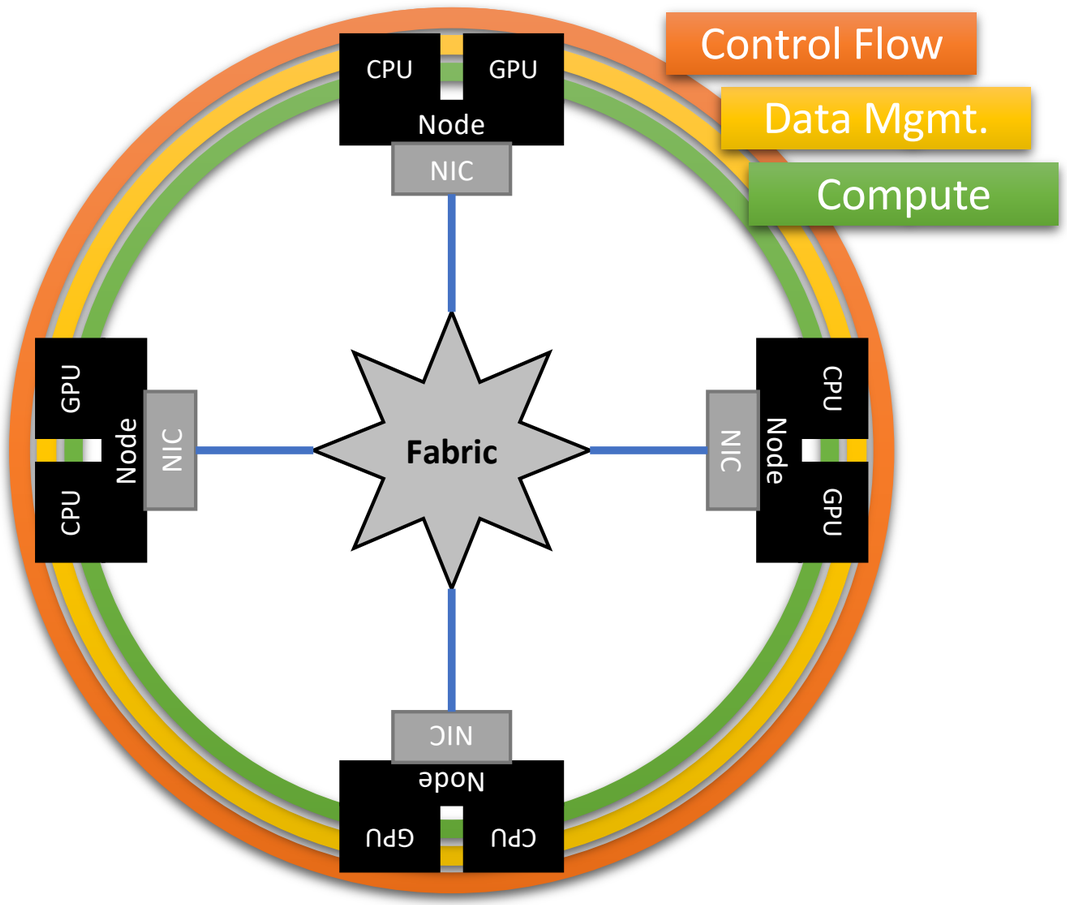
no fundamental change

- Cores on leave nodes still manage work and communication

Current State-of-the-Art

HPC systems are typically clusters

- Many equal, often fat nodes
- Mulit-core & accelerators
- Connected via passive networks



Consequence:

- Compute, Data Management and Control Flow only at core level

Disadvantages

- Scalability based on #cores
- Extra data transfers on local busses
- Network consumes power but does not contribute

ScalNEXT

SPONSORED BY THE
Federal Ministry of Education and Research

Project as part of German exascale software Program ScalExa

- October 2022-September 2025
- Partner: APS Network, JGU Mainz, LRZ Garching, KIT Karlsruhe, RWTH Aachen. TU Munich

Central goal: usage of smart networks to optimize HPC

- Development of interfaces and programming model extensions
- Across multiple application domains

Target platforms

- NVIDIA DPUs with ARM processors
- Silcom FPGA cards with Stratix-X
- APS Networks switch with Intel's Tofino

APS Networks JGU lrz KIT RWTH AACHEN UNIVERSITY TUM

The “Inverted” ScalNEXT Approach

Offload of control flow to SmartNIC

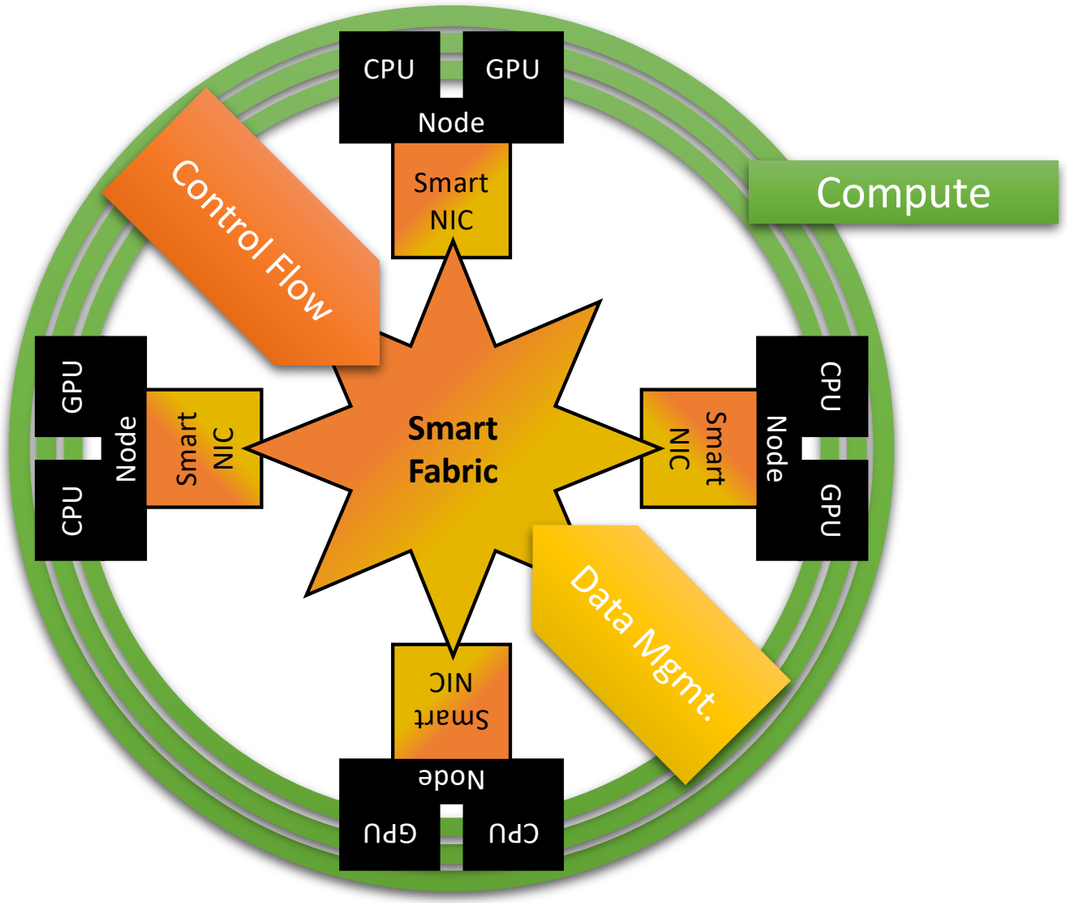
- Separate “compute” and “communicate”
- Reduction of number of end points
- Simplification of node architecture
- Usage of lean operating systems

Intelligent data processing in transit

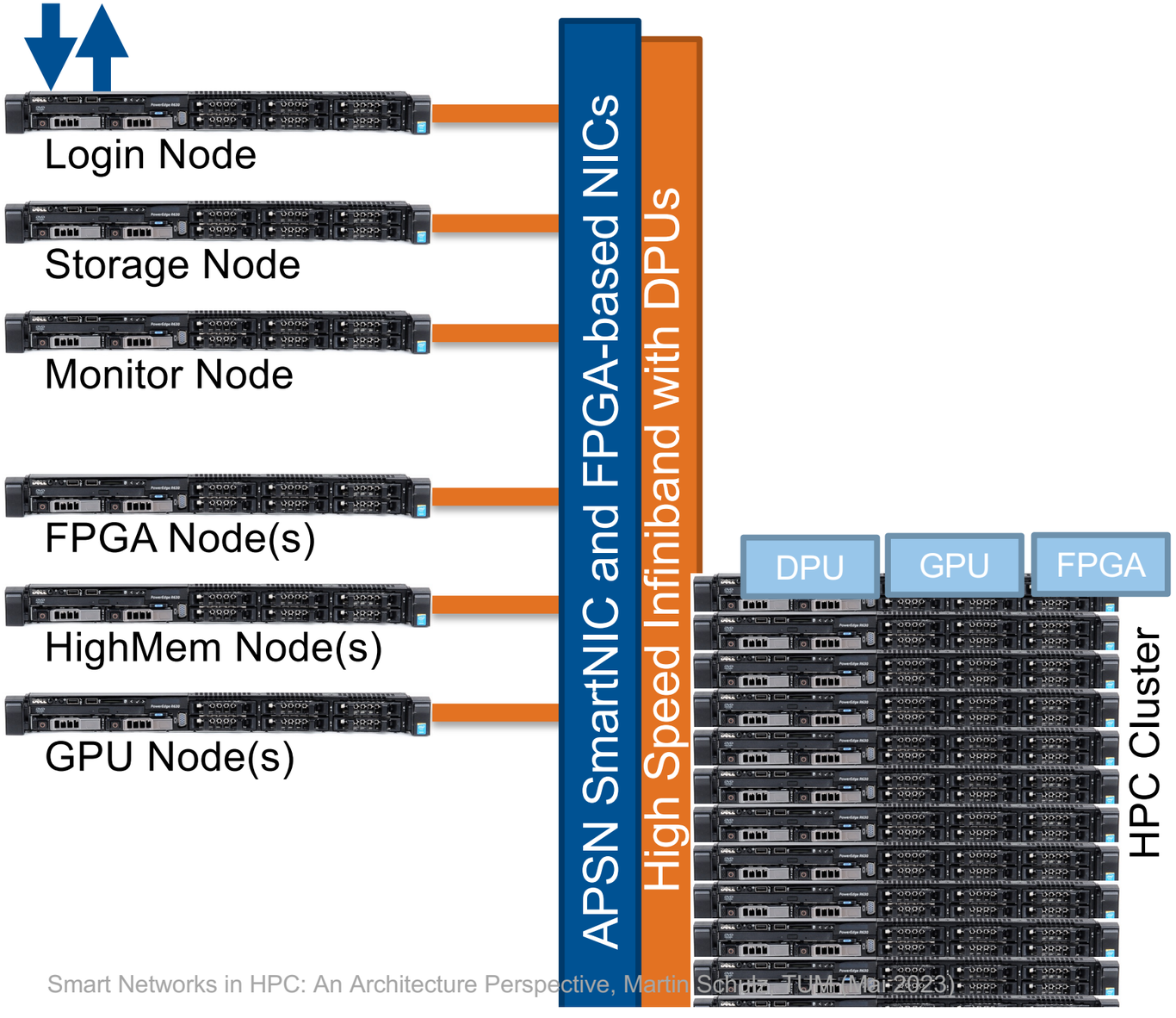
- Preprocessing near storage
- Data management in the fabric
- Optimized streaming

Usage of SmartSwitches

- Offload of “compute” to centrak resource
- Enable utilization of state in the fabric
- Utilize short communication distances

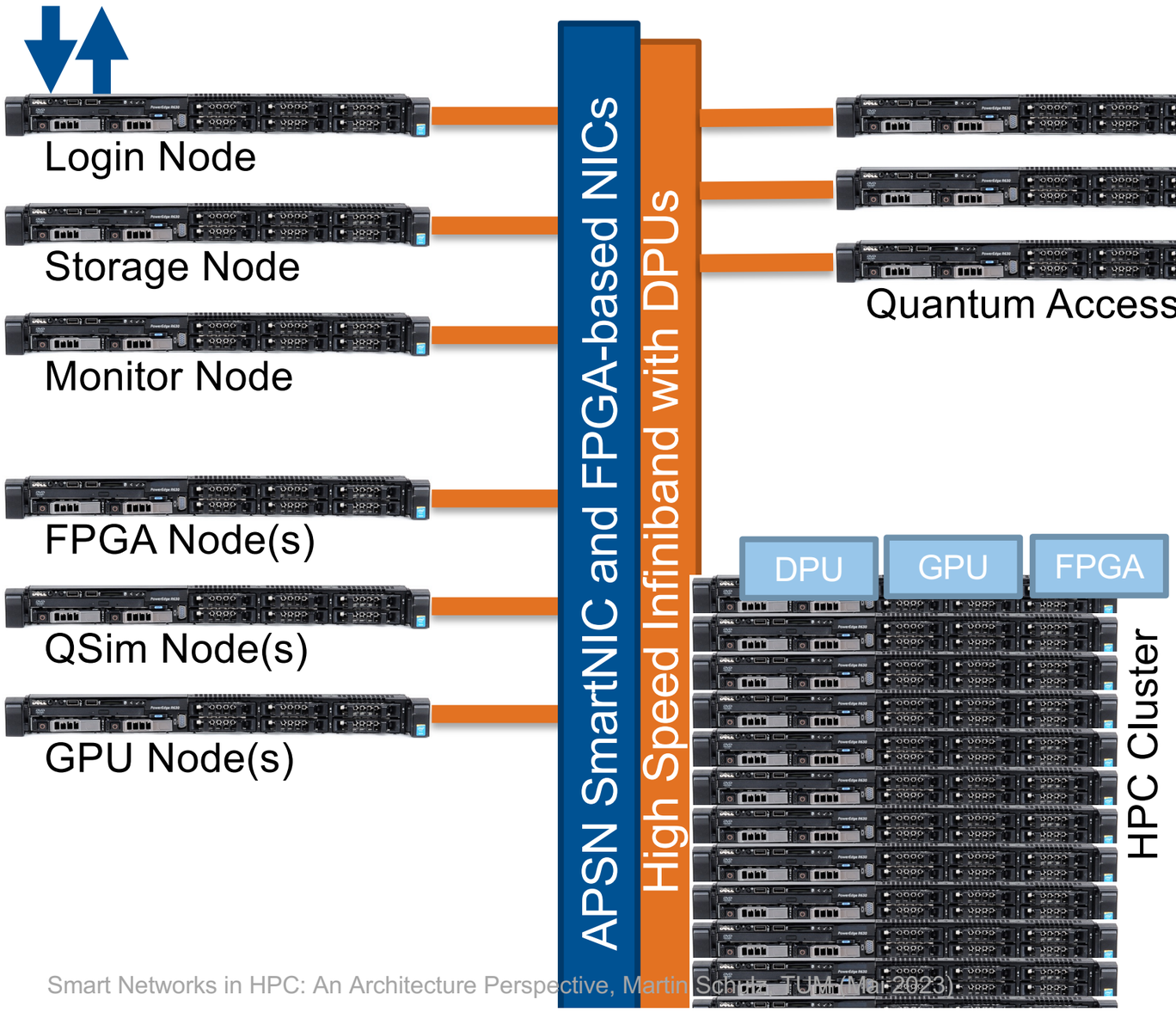


Wolpertinger System

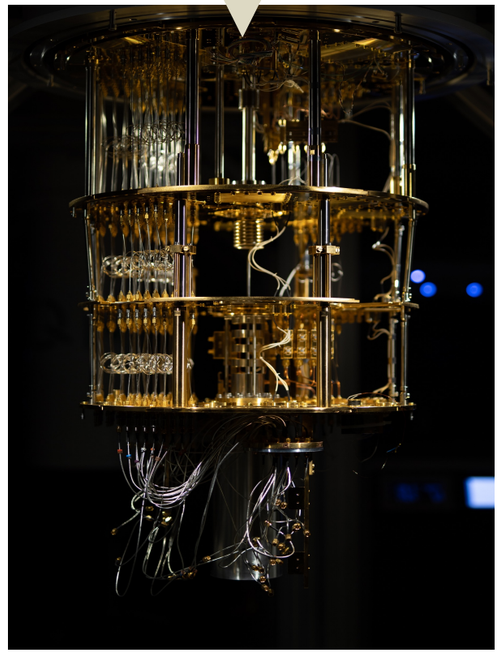


SPONSORED BY THE
Federal Ministry of Education and Research

Wolpertinger System

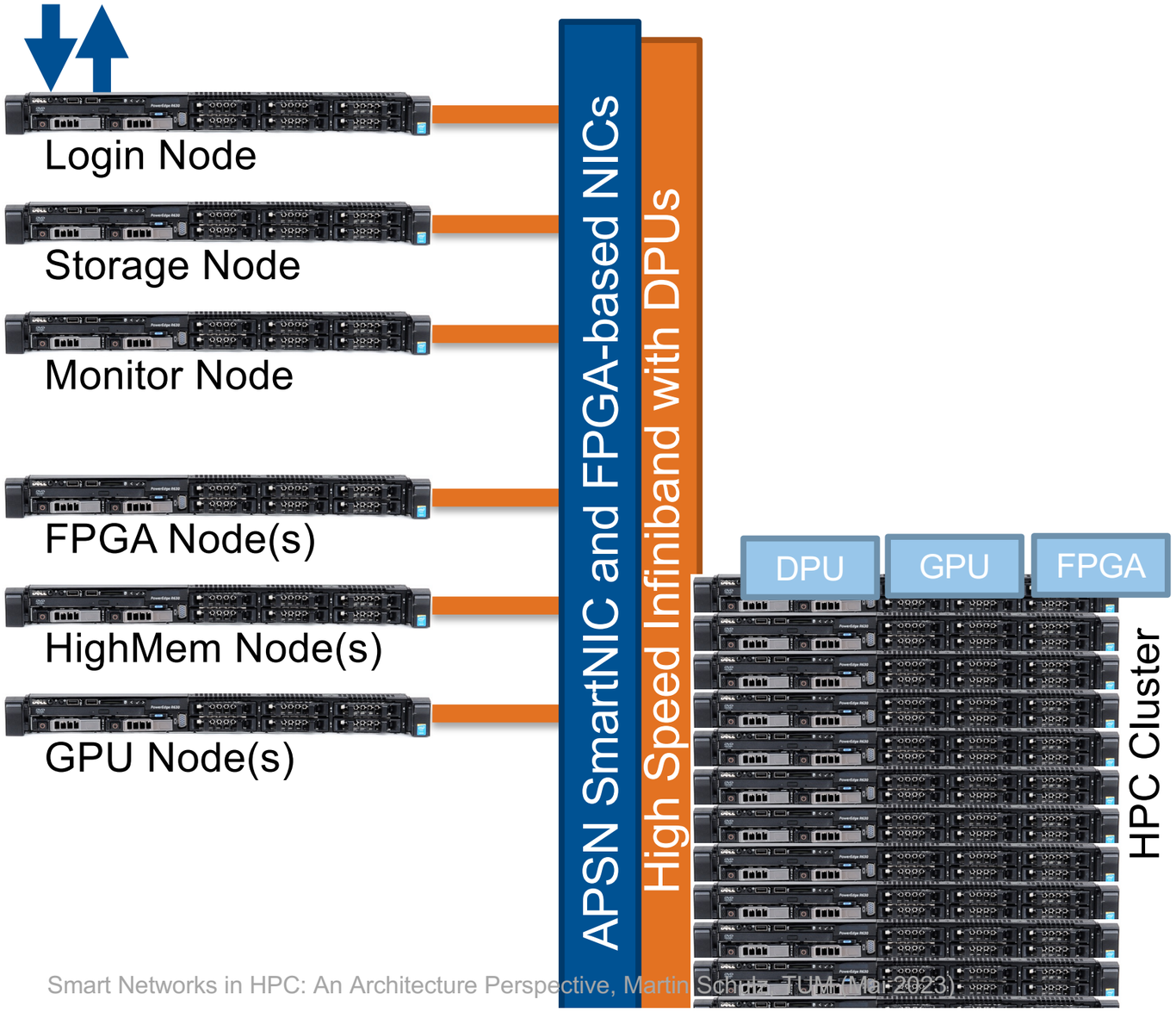


SPONSORED BY THE
Federal Ministry of Education and Research



Source: IQM Presskit

Wolpertinger System



SPONSORED BY THE

Federal Ministry of Education and Research

Resarch Environment

- Not a production system
- Flexible scheduling options
- Dedicated monitoring environment
- Ability to add more accelerators

Part of

- **B**avarian
- **E**nergy,
- **A**rchitecture &
- **S**oftware
- **T**estbed

Long Term Impact

Rethinking the node architecture

- Separate “compute” and “communicate”
- Tighte integration of compute, accelerator & network
- On-node tasking interfaces on top of
  → Hermit core (RWTH) & HalAdapt (KIT)

Designing Programming interfaces

- Separate specification of compute and communication
- Offload to and synchonization with NIC
- Reverse offload / onload opportunuties

Adding SmartNIC/Network support to MPI

- Hints and assertions on communicators/sessions
- Efficient merging of end-points into a single MPI process
- Specification of offload routines

Conclusions

SPONSORED BY THE
Federal Ministry of Education and Research

Smart networks can help with push past Exascale

- Going beyond standard offload of MPI functionality
- Inverted node architectures with increased scalability
- Near storage compute and in-network state

ScalNEXT project

- Explore SmartNICs (ARM&FPGA) and Switches
- New programming paradigms
- Across wide variety of application areas

Wishlist

- Extensive monitoring options (on and thru NIC)
- Software power control knobs
- Simplified host ←→ NIC interface with task synchronization
- Easy offload from host
- Vendor support via adoption of standards