OAK RIDGE National Laboratory

Computational Frontier: Heterogenous Computing

Travis Humble
Oak Ridge National Laboratory
humblets@ornl.gov

This presentation highlights advances in heterogenous computing capabilities for the Workshop on Experimental Applications of Artificial Intelligence for the Electron Ion Collider

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

This work is supported by the DOE Office of Science.

U.S. DEPARTMENT OF ENERGY

Modern High-Performance Computing

TITAN

Cray XK7, 18,688 Nodes
16-core AMD Interlagos + K20X
17 PFLOPS, 8.2 MW,
#1 TOP500 (2012)

SUMMIT

IBM, 4,600 Nodes
2 Power9 + 6 NVidia Volta
144 PFLOPS, 9.7 MW,
#1 TOP500 (2018)

FRONTIER

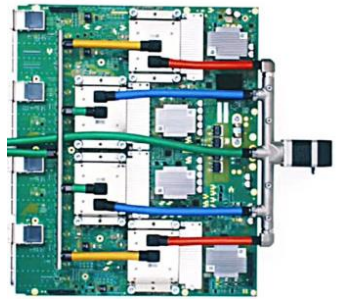
CRAY, 100 Cabinets
1 AMD EPYC + 4 Radeon Instinct
1.5 EXAFLOPS
Expected 2022

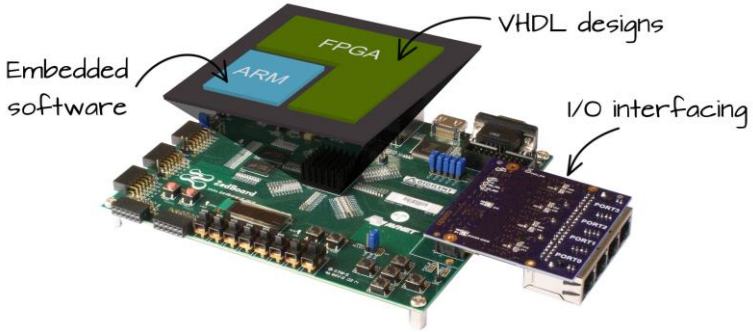
AND BEYOND

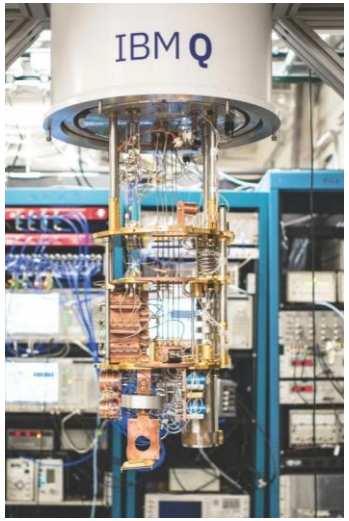
?

Scientific discovery and energy security depend on advances in computational capability

Modern Scientific Computing

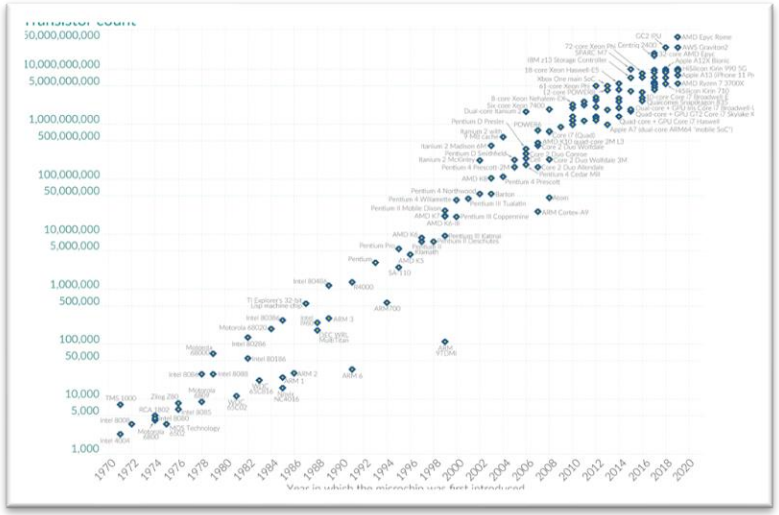


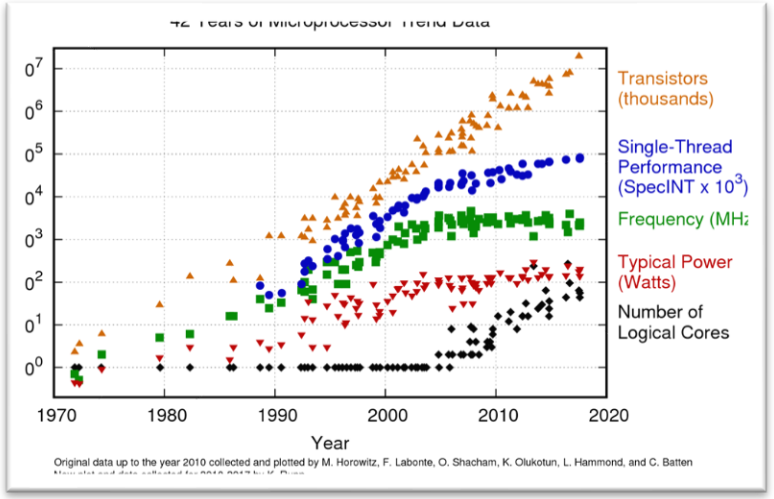




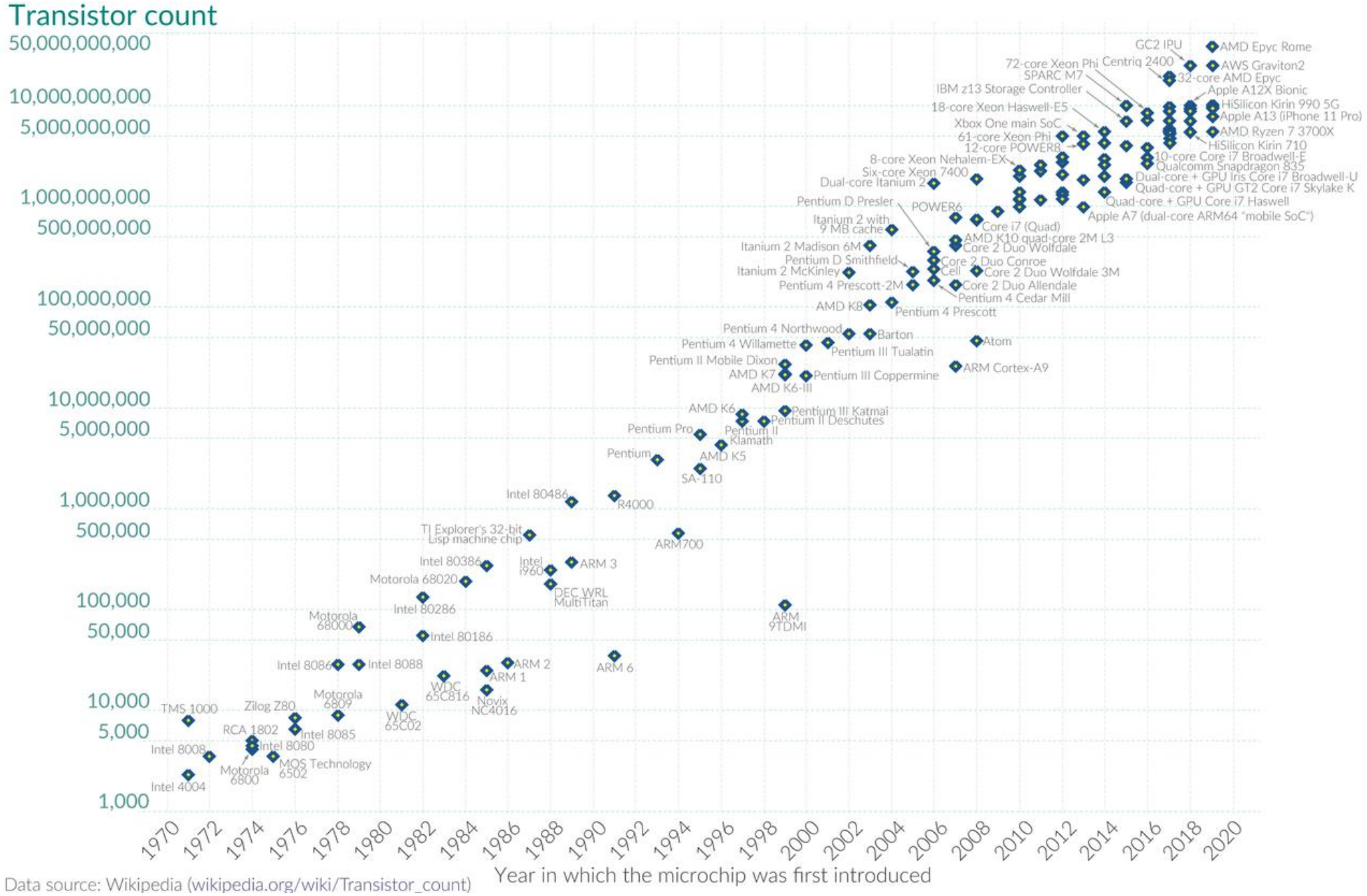
Heterogenous Computing

- In the past three decades, advances in computer technology have allowed the performance and functionality of processors to double every 2 years, i.e, Moore's law.
- The classical technological driver that has underpinned Moore's law for the past 50 years is already failing and is anticipated to flatten by 2025

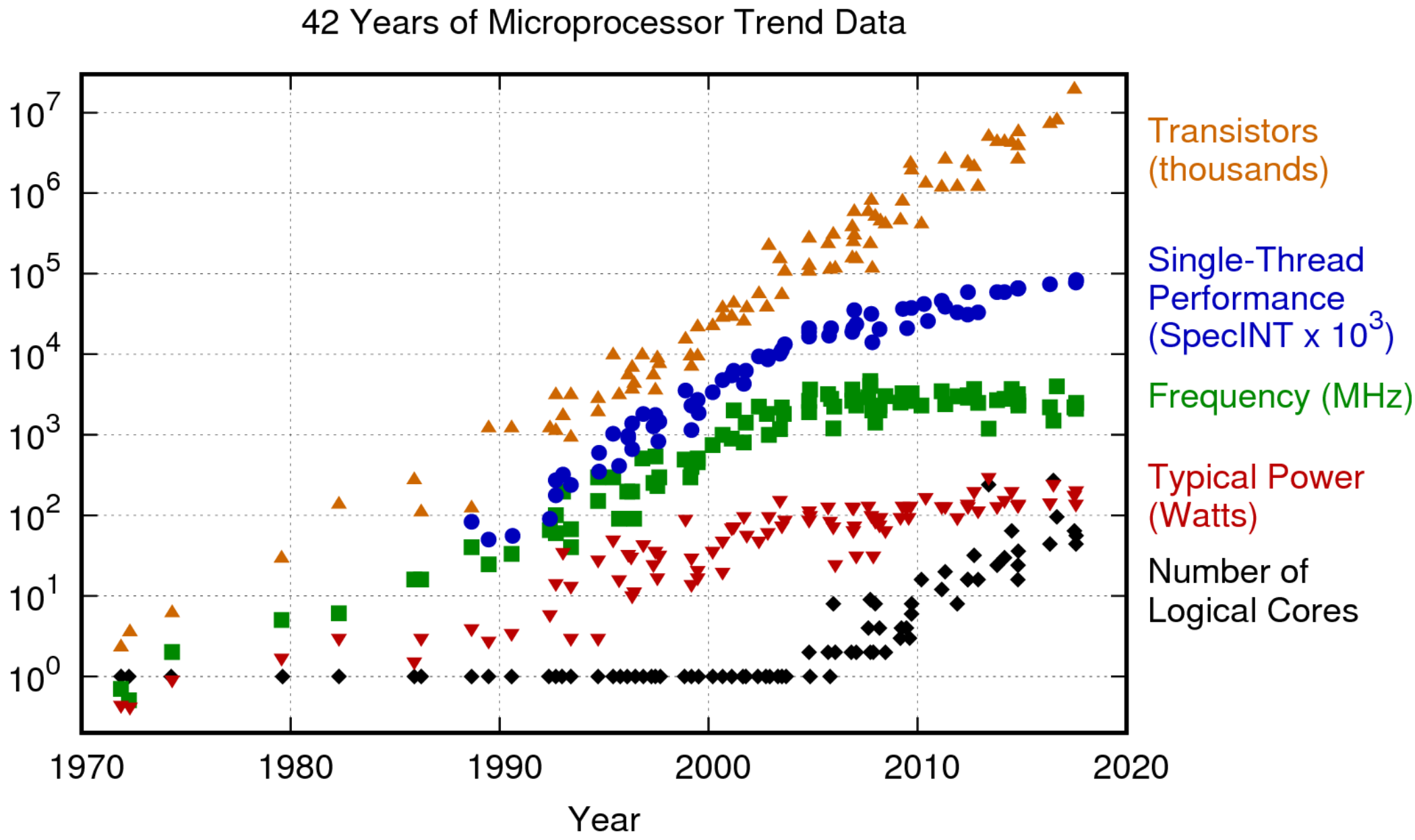




Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756



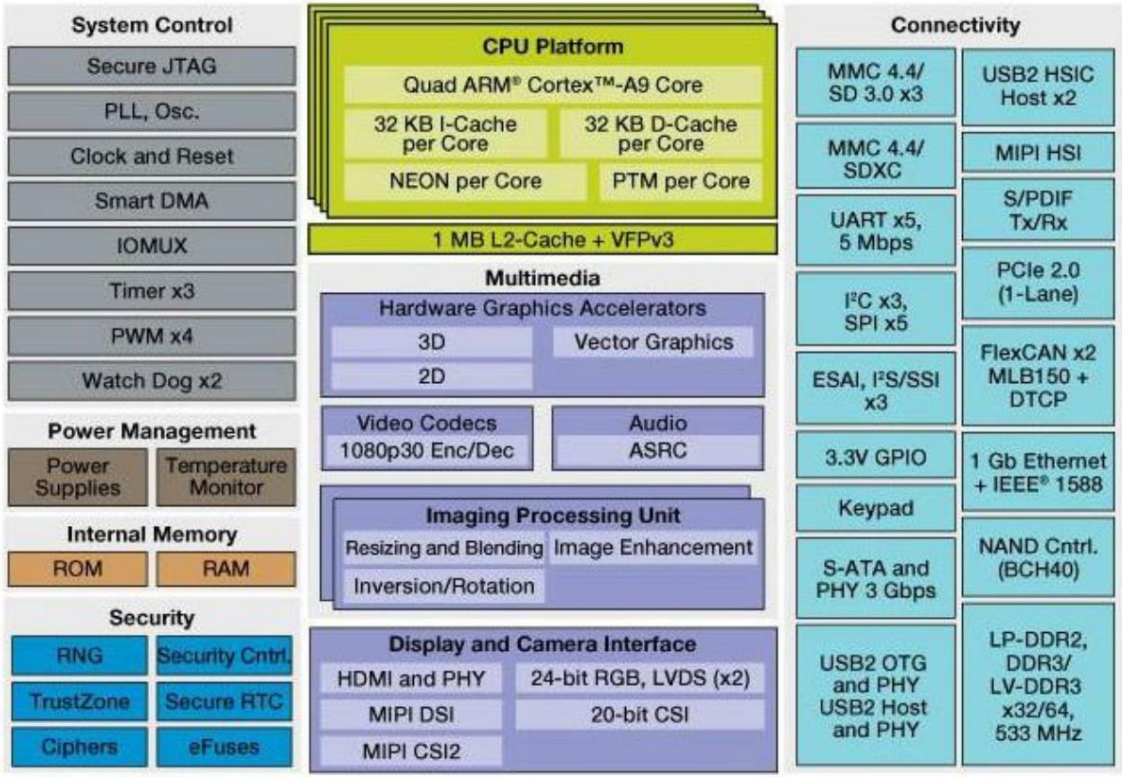
https://en.wikipedia.org/wiki/Transistor_count



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Heterogenous Computing

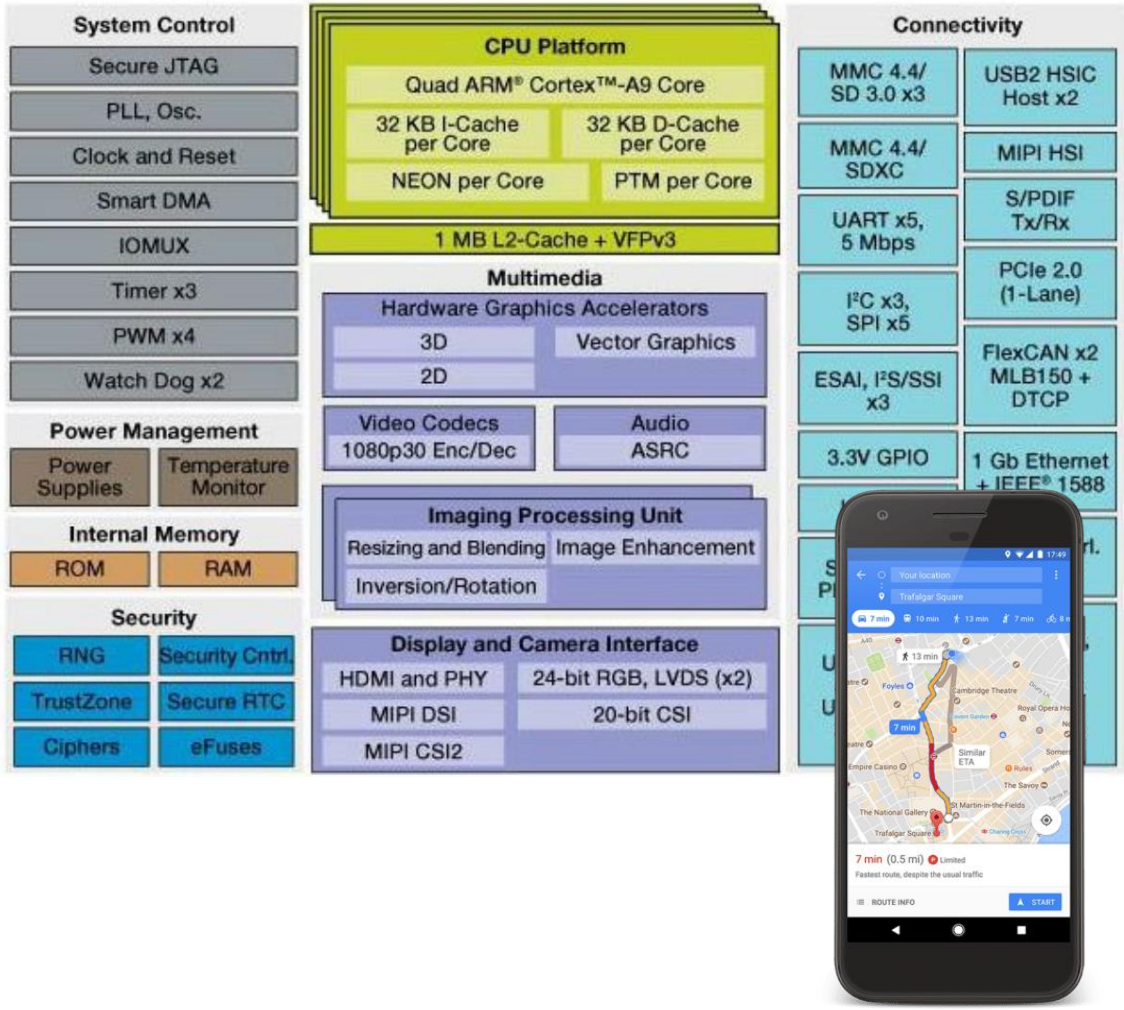
- The only tool left to a computer architect for extracting continued performance improvements is to use transistors more efficiently by specializing the architecture to the target scientific problem
- Computer vendors are pursuing systems built from combinations of different types of processors to improve capabilities, boost performance, and meet energy efficiency goals.



Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

Heterogenous Computing

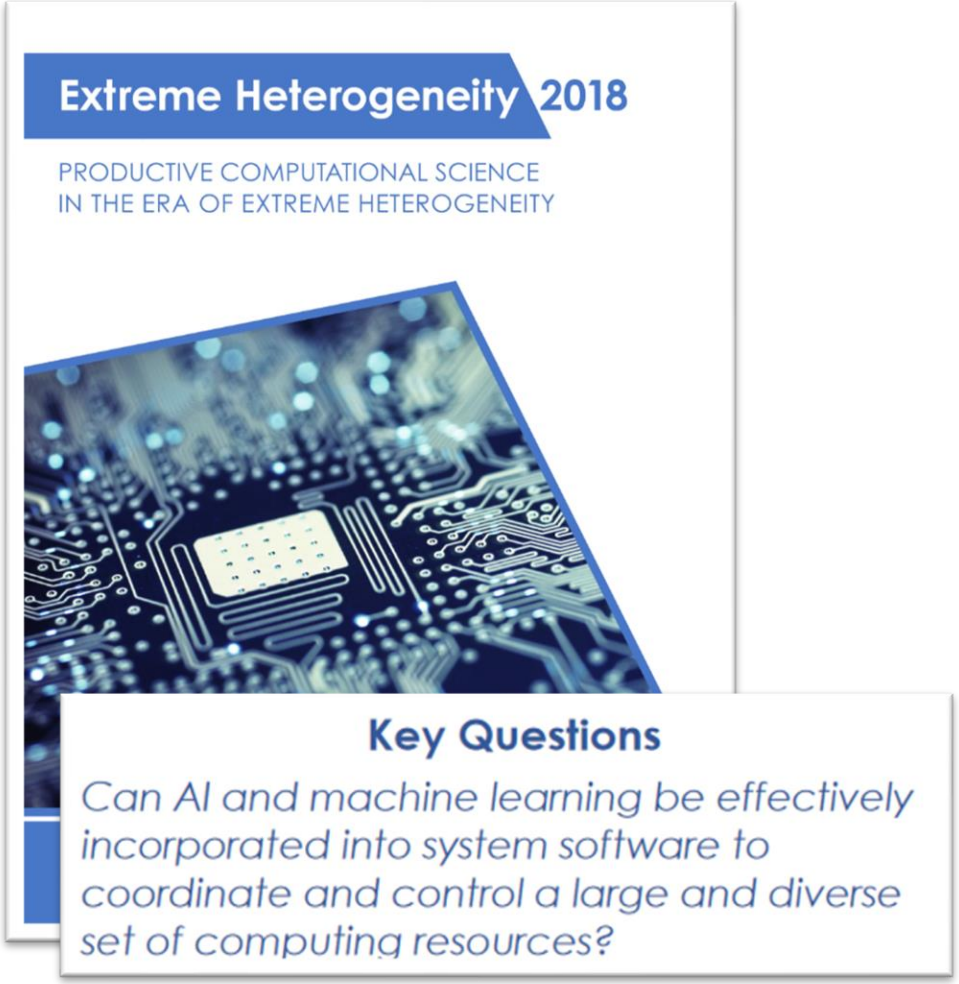
- The only tool left to a computer architect for extracting continued performance improvements is to use transistors more efficiently by specializing the architecture to the target scientific problem
- Computer vendors are pursuing systems built from combinations of different types of processors to improve capabilities, boost performance, and meet energy efficiency goals.



Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

Workshop on Productive Computational Science in the Era of Extreme Heterogeneity

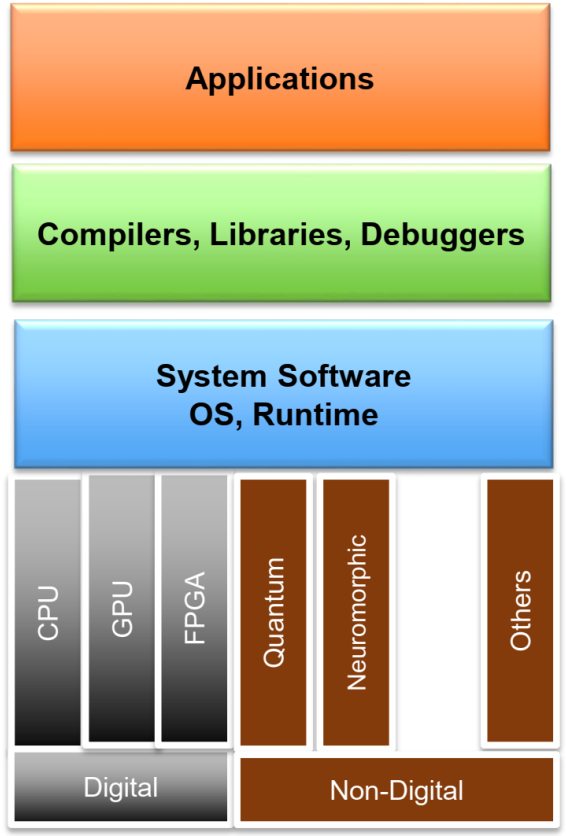
- Heterogeneity and diversity of architecture are nearly inevitable given current architecture trends
- External economic drivers and design diversity will result in systems built from a custom aggregation of components
- This fundamental change in computer architecture design has been deemed the era of "extreme heterogeneity."



Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

Computational Science in the Era of Heterogeneity

- The difficulty and complexity of developing scientific software will increase.
- New research in computer science is needed to make future supercomputers usable, useful and secure for science applications in the 2025-2040 timeframe



Areas of Key Finds

- Programming Environments
- Software Development, Sustainability, and Productivity
- Operating and Runtime Systems
- Modeling and Simulation
- Facilitating Data Management, Analytics, and Workflows

Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

Key Findings for Extreme Heterogeneity

- Programming Environments
  - Programmability
  - Mapping
  - Data Centricity
  - Correctness
- Software Development, Sustainability, and Productivity
  - Novel Design and Development Methodologies
  - Composability and Interoperability
  - Evolving Existing Software to New Programming Models
  - Reproducibility
- Operating and Runtime Systems
  - Design of Operating and Runtime Systems
  - Decentralized Resource Management
  - Autonomous Resource Optimization
- Modeling and Simulation
  - Methodology and Tools for Accurate Modeling of Extreme Heterogeneity
  - Cohesive Integration of Modeling and Simulation Infrastructure with Programming Environment and Runtime System
- Facilitating Data Management, Analytics, and Workflows
  - Mapping Science Workflows to Heterogeneous Hardware and Software Services
  - Adapting Workflows and Services Through Learning Approaches

Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

Five Priority Research Directions

1. Maintaining and Improving Programmer Productivity

- *Flexible, expressive, programming models and languages*
- *Intelligent, domain-aware compilers and software development tools*
- *Composition of disparate software component content*

2. Managing System Resources Intelligently

- *Automated methods using introspection and machine learning*
- *Optimize for performance, energy efficiency, and availability*

3. Modeling and Predicting Performance

- *Evaluate impact of potential system designs and application mappings*
- *Model-automated optimization of applications*

4. Enabling Reproducible Science Despite Diverse Processors and Non-Determinism

- *Methods for validation on non-deterministic architectures*
- *Detection and mitigation of pervasive faults and errors*

5. Facilitating Data Management, Analytics, and Workflows

- *Mapping a science workflow to heterogeneous hardware and software services*
- *Adapting workflows and services through machine learning approaches*

Productive Computational Science in the Era of Extreme Heterogeneity https://doi.org/10.2172/1473756

AI Chip Landscape

https://basicmi.github.io/AI-Chip/

▪ MLPerf results available ▪ AI-Benchmark results available

Tech Giants/System

- Google ▪ TPU ▪ Edge TPU
- Microsoft ▪ Brainwave
- facebook
- aws ▪ Inferentia
- Apple ▪ A13
- IBM
- Alibaba Group 阿里巴巴集团 ▪ Hanguang800 ▪ TG6100N
- Huawei HiSilicon ▪ Ascend ▪ Kirin ▪ Hi35xx
- Baidu 百度 ▪ Kunlun/ Honghu
- Tesla ▪ FSD
- Tencent 腾讯
- Hewlett Packard Enterprise
- Fujitsu ▪ DLU
- DELL
- inspur 浪潮
- Western Digital
- NOKIA
- LG

IC Vendor/Fabless

- Intel ▪ NNP-I ▪ NNP-T/Myriad X/ EyeQ/Arria FPGA
- Samsung ▪ Exynos 9825
- NVIDIA ▪ Volta/Turing /T4/Xavier/NVDLA
- Qualcomm ▪ Snapdragon 855 ▪ Cloud AI 100
- AMD ▪ EPYC
- Xilinx ▪ VERSAL
- MediaTek ▪ Dimensity
- Unisoc ▪ Tiger T710
- Marvell
- Rockchip 瑞芯微电子 ▪ RK3399Pro
- Ambarella ▪ CV22S/25S
- NationalChip ▪ GX8010

Automated Driving

- NXP
- Texas Instruments
- Renesas
- Toshiba
- ST

Startup in China

- Cambricon 寒武纪科技 ▪ MLU100/270/220
- 地平线 Horizon Robotics ▪ Journey
- Bitmain ▪ BM1682/1880
- intelliFusion 云天励飞 ▪ DeepEye1000
- 依图 YITU ▪ QuestCore
- 肇观电子 NextVPU ▪ N171
- Canaan ▪ K210
- artosyn 酷芯微电子 ▪ AR9000
- 深鉴科技 Intengine ▪ Voitist611
- Tsing Micro ▪ TX101/210/510
- 亿智科技 ▪ TAi8010
- Black Sesame Technologies ▪ HuaShan
- Enflame ▪ DTU
- Witmem 知存科技 ▪ MemCore001

Smart Voice

- Chipintelli ▪ CI100X/110X
- ▪ CSK4002
- 云知声 Unisound
- AISpeech 思必驰 专注人性化的智能语音

Startup Worldwide

- Cerebras ▪ WSE
- Wave Computing
- Graphcore ▪ GC2
- SambaNova Systems
- Habana ▪ Gaudi ▪ Goya
- Hailo ▪ Hailo-8
- Blaize Xplorer X1000
- Kalray ▪ MPPA2-256
- Groq
- Tachyum
- Esperanto Technologies
- PEZY Computing ▪ PEZY-SC2
- Eta Compute ▪ ECM3531
- GreenWaves Technologies ▪ GAP8

FPGA/eFPGA

- Achronix
- flexlogix
- EFINIX

Processing in Memory

- Mythic
- Syntiant
- Areanna
- gyrfalcon technology ▪ GAINBOARD/ Lightspeeur
- Anaflash

Optical Computing

- Lightelligence
- Lightmatter
- Optalysys
- Luminous

Neuromorphic

- brainchip
- aiCTX
- Rain Neuromorphics
- GML Grid Matter Labs
- ABR

- Preferred Networks ▪ MN-Core
- Innogrit ▪ Shasta/Rainier/Tacoma
- Kneron ▪ KL520
- Aistorm
- NeuroBlade
- Novumind
- Tenstorrent
- Aimotive

More at https://basicmi.github.io/AI-Chip/

IP/Design Sevice

- arm
- Synopsys
- Imagination
- CEVA
- cadence
- SiFive
- Arteris IP
- Moortec

Design service with In-house IP

- VeriSilicon
- Broadcom
- GUC
- alchip
- Faraday
- eSilicon

Compilers

- TensorFlow | XLA | MLIR
- Glow
- tvm
- OctoML
- NVIDIA | TensorRT
- PlaidML
- nGraph
- ONNC
- Tiramisu Compiler
- The Tensor Algebra Compiler (taco)

Benchmarks

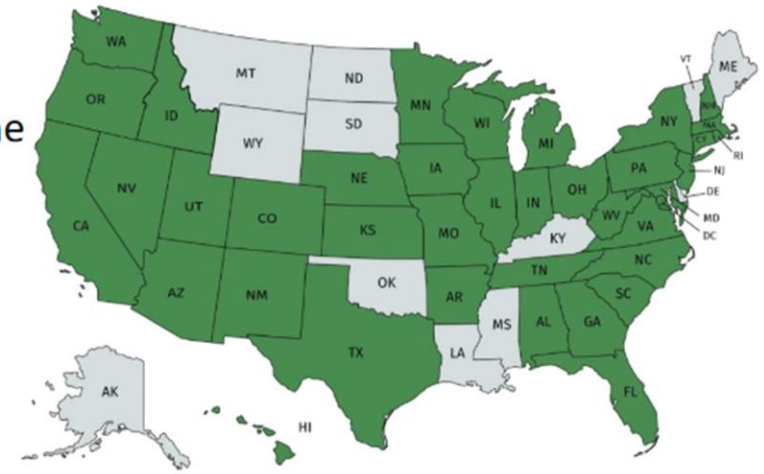
- MLPerf
- AI - Benchmark
- AI Matrix.
- 中国人工智能产业发展联盟 Artificial Intelligence Industry Alliance
- DAWNBench
- MLMark An EEMBC Benchmark

Artificial Intelligence for Science

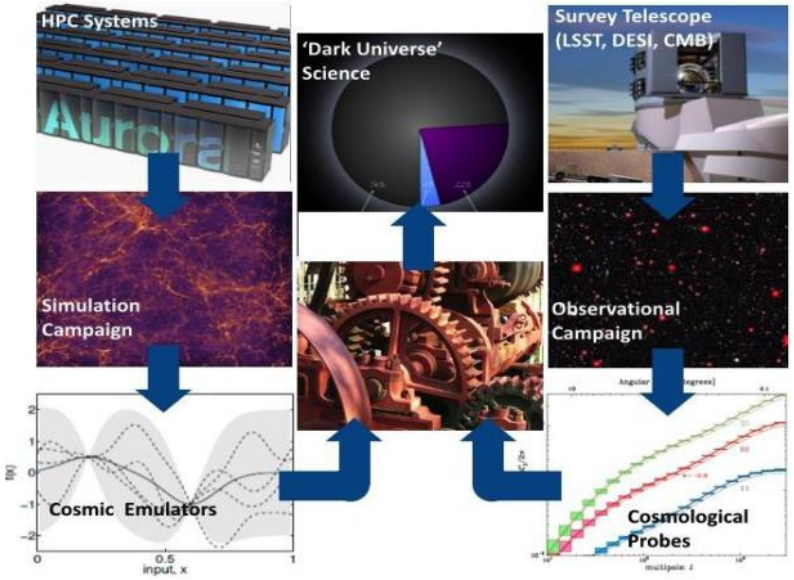
The AI for Science Town Halls so far

- Over 1000 registrations across 4 Town Halls

ANL	July 22-23	357	
ORNL	Aug 20-21	330	
LBNL	Sept 11-12	349	+100 online
DC	Oct 22-23	273	+ ?
Totals		1309	1309



- All 17 DOE National Laboratories
- 39 Companies from large and small
- Over 90 different universities
- 6 DOE/SC Offices + EERE and NNSA



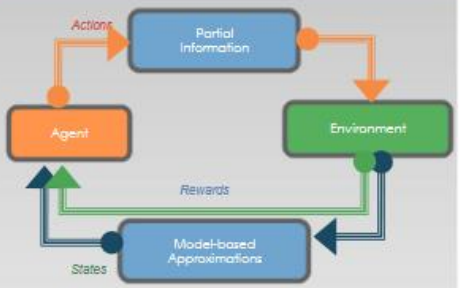
AI for Science https://doi.org/10.2172/1604756

Research Challenges in Heterogenous AI

Data

- Experimental design
- Data curation and validation
- Compressed sensing
- Facilities operation and control

Learning

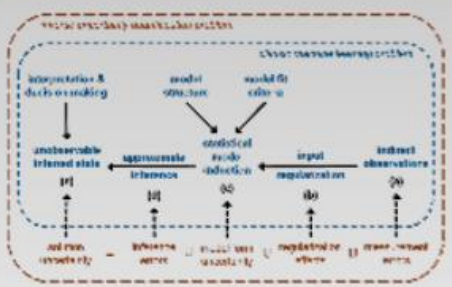


- Physics informed
- Reinforcement learning
- Adversarial networks
- Representation learning and multi-modal data

Scalability

- Algorithms, complexity and convergence
- Levels of parallelization
- Mixed precision arithmetic
- Communication

Assurance



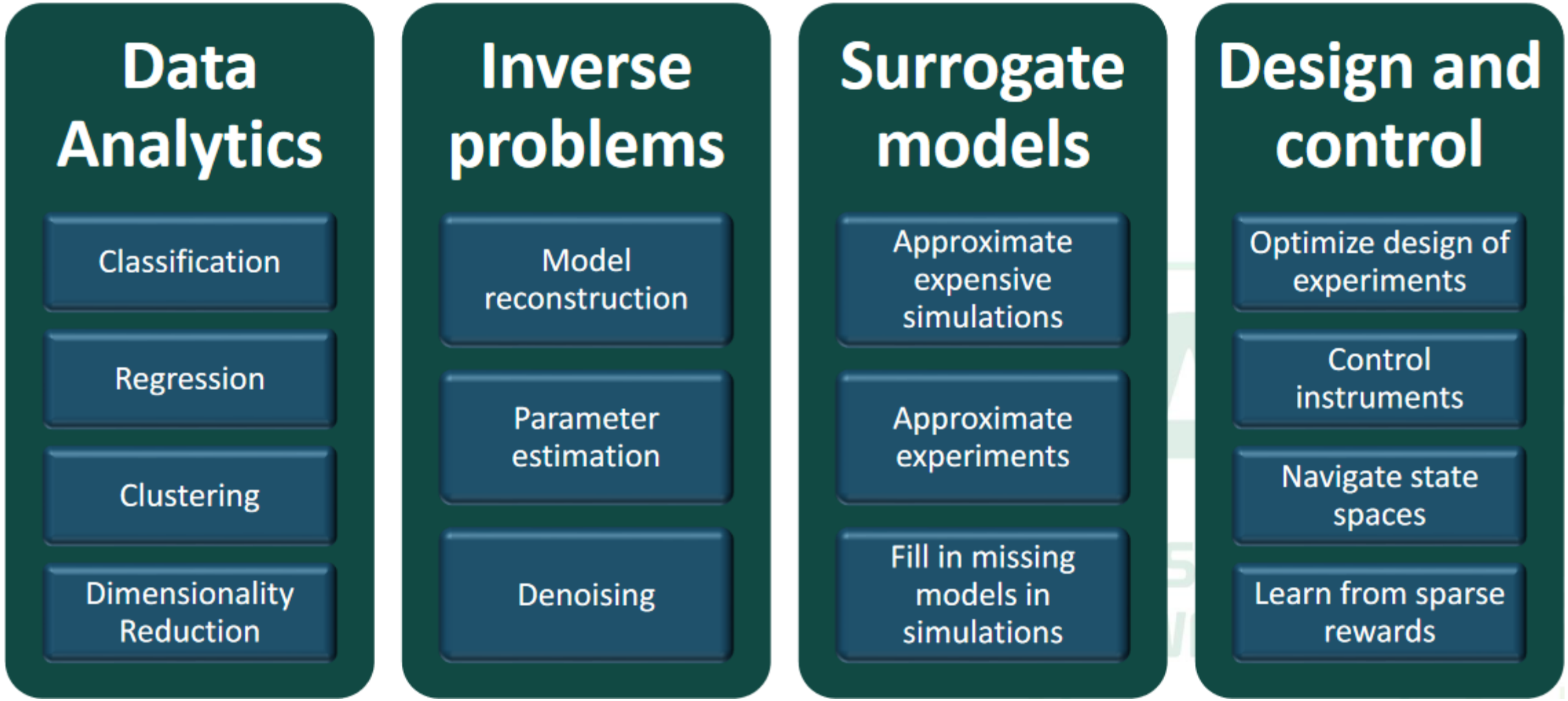
- Uncertainty quantification
- Explainability and interpretability
- Validation and verification
- Causal inference

Workflow

- Edge computing
- Compression
- Online learning
- Federated learning
- Infrastructure
- Augmented intelligence

AI for Science https://doi.org/10.2172/1604756

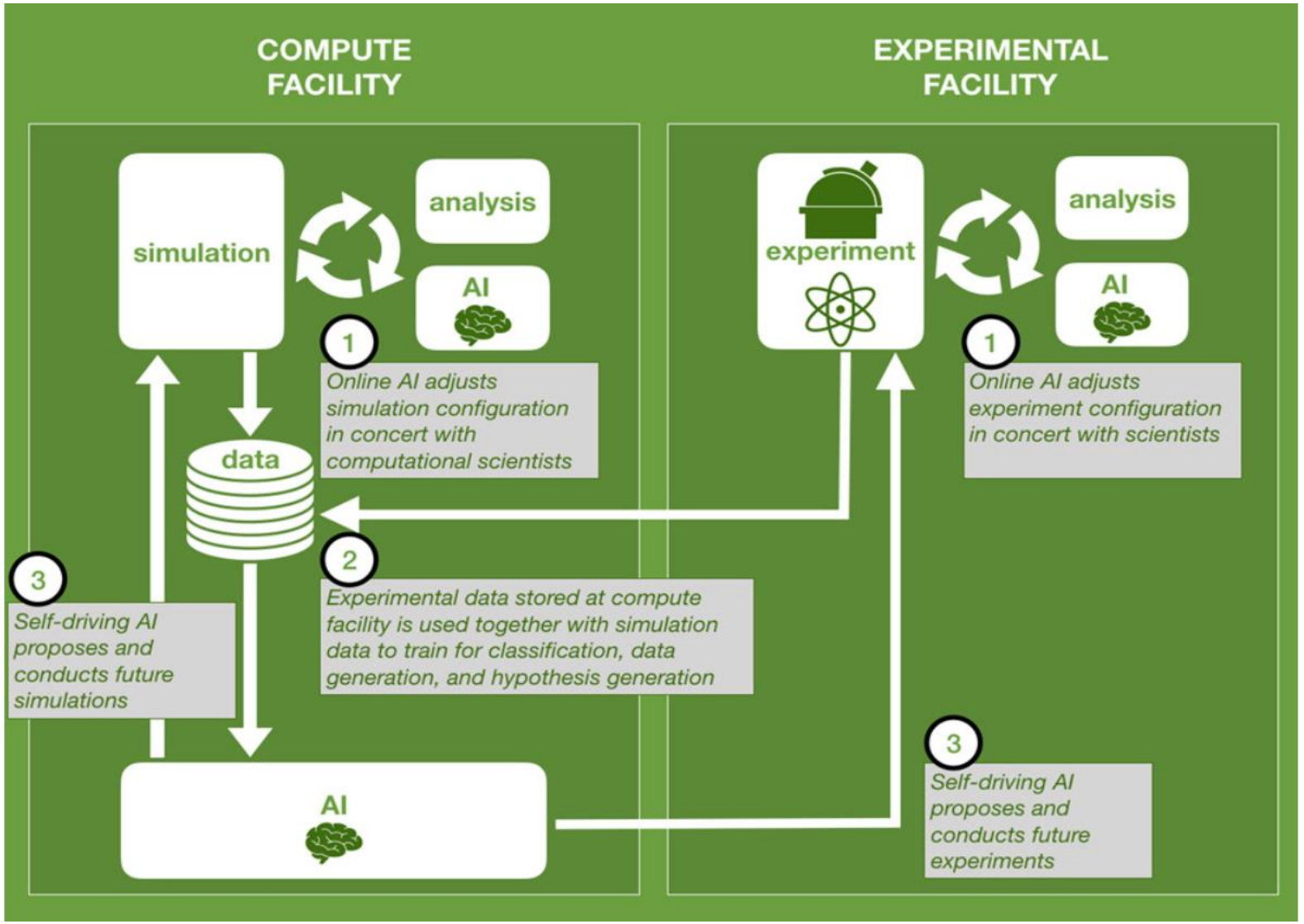
Opportunities for Artificial Intelligence



AI for Science https://doi.org/10.2172/1604756

Opportunities for Artificial Intelligence

1. The "inner loop" of simulations and experiments
2. Accelerate and enhance traditional analysis approaches
3. The "outer loop" to assist in the management and control of workflows, laboratories, and facilities.



AI for Science https://doi.org/10.2172/1604756

Computational Frontier: Heterogenous Computing

- Heterogeneity and diversity of architecture are nearly inevitable given current architecture trends
- The difficulty and complexity of developing scientific software will increase.
- The opportunities of AI for Science are spurred by heterogeneous computing

Priority Research Directions

- Maintaining and Improving Programmer Productivity
- Managing System Resources Intelligently
- Modeling and Predicting Performance
- Enabling Reproducible Science Despite Diverse Processors and Non-Determinism
- Facilitating Data Management, Analytics, and Workflows