ACCELERATING COMPUTING FROM THE EDGE TO THE DATACENTER IN THE NEXT DECADE
TRADITIONAL HPC WORKLOAD
50 Years of Microprocessor Trend Data

Transistors (thousands)

Single-Thread Performance (SpecINT x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

Year


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 2010-2021 by K. Rupp
DENNARD SCALING ENDS

50 Years of Microprocessor Trend Data

Transistors (thousands)
Single-Thread Performance (SpecINT x 10^3)
Frequency (MHz)
Typical Power (Watts)
Number of Logical Cores

Year

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 2010-2021 by K. Rupp
I Wrote My First Simulation

DENNARD SCALING ENDS

50 Years of Microprocessor Trend Data

Transistors (thousands)

Single-Thread Performance (SpecInt x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

Year


50 Years of Microprocessor Trend Data

- Transistors (thousands)
- Single-Thread Performance ($\text{SpecINT} \times 10^3$)
- Frequency (MHz)
- Typical Power (Watts)
- Number of Logical Cores

DENNARD SCALING ENDS

HIGGS Boson Discovered

ImageNet Victory

50 Years of Microprocessor Trend Data

Transistors (thousands)

Single-Thread Performance (SpecINT x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

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 2010-2021 by K. Rupp
EXPANDING UNIVERSE OF THE NEW AGE OF HPC

EDGE
HPC * AI
SIMULATION
DIGITAL TWIN
QUANTUM COMPUTING

SUPERCOMPUTING

EDGE
NETWORK
CLOUD
EXTREME IO
NEW WORKFLOWS EMERGING TO SOLVE GRAND CHALLENGES

- HPC*AI
- DIGITAL TWINS
- EXPERIMENTS /SENSORS
- QUANTUM COMPUTING
- SIMULATION

Quantum circuit simulation
Earth 2
Turbine wake optimization
NEW WORKFLOWS EMERGING TO SOLVE GRAND CHALLENGES
Integrated workflow with real-time analysis, steering and visualization for human in the loop
Optimize and bring the best solution to that create the integrated workflow
Optimize and bring the best solution to that create the integrated workflow

Workstreams

Instrument to Edge
Data Transmission
Sim + DM/A + AI

Workflow between Datacenter & Experiment

Control System for Experiment
Sensor/Instrument
Data Acquisition
Inference
Data Analytics Post Process
Results

Sim + DM/A + AI

Optimize and bring the best solution to that create the integrated workflow

NV SW

RAPIDS
Triton/TensorRT
Holoscan, Morpheus, Issac, UCF

Aerial
DOCA
Morpheus

HPC SDK, RAPIDS, DL FW,
Omniverse, Modulus, UC

NV HW

Jetson, AGX, EGX
DGX Station/Server

DPU
MetroX

DGX Server/SuperPOD
HGX
Build their applications using a mix of C++, Python, JAX

Develop AI microservices combining low-latency data streaming while passing more complex tasks to data center resources

Scale from embedded to datacenter
ADVANCED LIGHT SOURCE : @ LBNL

From ~105 seconds to 12 seconds

- For more details watch: Accelerating Sensor Processing Pipelines with NVIDIA Toolkits
- The GTC talk may reference internal names used during initial development
UKAE EARLY VERSION OF DIGITAL TWIN FOR TOKAMAK REACTOR
ENGAGEMENTS WITH HPC*AI*EDGE

Workflow between Datacenter & Experiment

Control System for Experiment Sensor/Instrument Data Acquisition Inference Data Analytics Post Process Results

Sim + DM/A + AI

Workflows

Instrument to Edge  Data Transmission  Sim + DM/A + AI

Workstreams

ALS/LBNL  CNMS/ORNL  APS/ANL  DIII-D/GA/UKAE  LHCb/CERN  HED Physics/LLNL

Optimizing Ptychography pipeline Automating Microscopy AI accelerated Nanoscale x-ray imaging AI surrogate, CGYRO, Digital Twin Design complete for using NVIDIA A40 for HLT HRR HED laser plasma experiments workflow
SUPERCHARGING SCIENCE EXPERIMENTS AND INSTRUMENTS

ANL/APS ACCELERATES X-RAY PTYCHOGRAPHY 300X WITH PTYCHONN

ADVANCED BIOIMAGING CENTER @UC-BERKELEY REAL TIME LIVE CELL IMAGING LIGHT SHEET MICROSCOPY

PtychoNN paper: AI-enabled high-resolution scanning coherent diffraction imaging

Link to keynote video - https://youtu.be/rXG27G3bWzY
RISE OF HPC AT THE EDGE
Posing a New Set of Challenges for HPC

10X - 100X MORE DATA
50+ GIANT SCALE INSTRUMENTS WW

ELT ESO
LIGO
SKA

AI SUPERCOMPUTING AT THE EDGE
ENABLES REAL-TIME INSIGHTS AND CONTROL

100x Data Collected
1x Data Transfer

Real-Time Streaming AI Processing

ML/AI Model

Edge __________ Datacenter

STREAMING DEPLOYMENT IS HARD
FOR DATA SCIENTISTS, RESEARCHERS AND DEVOPS

Streaming Data Performance
Easily Scale Implementation
Developer Ease-of Use
Combining multiple datastreams
ASSIMILATION OF SENSOR DATA ACROSS MULTIPLE INDUSTRIES

VISION

Harmonize the streaming AI framework architecture for developing cloud native, disaggregated scalable applications from embedded systems to Datacenter

Maximize reuse

Modular

Jetson Appliances  AGX  EGX/OEM Servers  DGX/OEM Servers

NVIDIA Confidential. For Use under NDA with ORNL
COMPOSING AN HPC STREAMING DATA PIPELINE USING STREAMING REACTIVE FRAMEWORK (SRF*)

SRF* is a reactive, network-aware, flexible, and performance-oriented streaming data framework that standardizes building modular and reusable pipeline mixing C++, Python, JAX.

- Asynchronous computation and mitigation of I/O and GPU blocking
- Distributed computation with message transfers over RMDA using UCX
- Dynamic reconfiguration to scale up and out at runtime
- Designed to mitigate backpressure with concurrent blocking queues between stages
- Hybrid HPC and Cloud Native

*SRF is under development. Final name subject to change
ANATOMY OF A SRF PIPELINE

A SRF pipeline is composed of **Segments**

- **Segments** are composed of **Sources**, **Sinks**, and **Nodes** (Source + Sink)
- **Segments** also guaranteed compute within a single node, can connect nodes via network (Edge, Cloud, or Datacenter), and contain MPI support
- **Nodes** process an input stream, create an output stream, and can be implemented with Python or C++
- Components are linked by **Edges** which are implemented as **Channels**
- **Channels** move data from sources to sinks and provide a backpressure policy
TO LEARN MORE ABOUT SENSOR DATA PROCESSING

- Blog [Facing the Edge Data Challenge with HPC + AI](#)
- GTC Spring 2022 [High Performance Geospatial Image Processing at the Edge*](#)
  - Geospatial image analysis using DPUs in an edge device designed to meet the Size-Weight-and-Power requirements for aircraft deployment.
- PtychoNN paper: [AI-enabled high-resolution scanning coherent diffraction imaging](#)
  - The Advanced Photon Source at Argonne National Laboratory runs PtychoNN on an Orin AGX at the x-ray detector. It is available for use at other light sources around the world.
- GTC Spring 2022 [Accelerating Sensor Processing Pipelines with NVIDIA Toolkits*](#)
  - Faster imaging pipelines by using JAX and SRF to processing streaming data with applications in Ptychography and Microscopy
- See the SRF description above and the [GitHub](#) page

*The GTC talk may reference SRF as ”Neo” which was the internal name used during initial development*
DIGITAL TWINS
DIGITAL TWIN: ACTIONABLE RESULTS AN ACTIONABLE TIME
DIGITAL TWINS WILL EXIST AT EVERY SCALE
DIGITAL TWIN AT REALISTIC COMPLEXITY

- Physical System
  - Virtual Sensor
  - Physics-based Models
    - ...
  - Physics-based Models

- Source of Truth
  - Human in the Loop
  - Digital Twin
    - Data Harvesting
    - (AI) Surrogate
      - ...
    - (AI) Surrogate
    - Virtual Actor
OMNIVERSE: PLATFORM FOR BUILDING DIGITAL TWINS

Physical System

Virtual Sensor

Physics-based Models

... Physics-based Models

Digital Twin

Data Harvesting

(AI) Surrogate ...

(AI) Surrogate

Virtual Actor

Human in the Loop
EXAMPLE: BEAM PATTERN EXPLORATION FOR PLACING 5G ANTENNA
ADVANCED TOOLS AND TECHNOLOGIES
Foundational Platform Components

- **NUCLEUS**: Source of truth
- **CONNECT**: Coupling
- **KIT**: Application API
  User experience
- **SIMULATION**: Virtual Actor
- **RTX RENDERER**: Virtual Sensor
DATA HOMOGENIZATION VIA USD

UNIVERSAL SCENE DESCRIPTION
The “HTML” of 3D Virtual Worlds

- Developed by Pixar
- Foundation for NVIDIA Omniverse
- Open-sourced API and file framework for complex scene graphs
- Easily extensible, simplifies interchange of assets between industry software
- Introduces novel concept of layering
- Enables simultaneous collaboration for large teams in different department working on the same scene
- Originated in M&E, now becoming a standard across industries including AEC, Manufacturing, Product Design, Robotics
OMNIVERSE NUCLEUS
Asset Database and Collaboration Engine

- Allows multiple software tools to talk to each other as well as live sync workflow
- Universal asset exchange - can house assets of any filetype
- Enables collaboration on large, ultra-complex scenes and passes only the change deltas
- Because only deltas are exchanged, extremely fast creation/replication is enabled
- No more hour-long or overnight uploading/downloading of entire scene files - everything is real-time and live
- Enables a single source of truth and eliminates messy, redundant file copies
• Integration of open source science application (GEANT4)
• FAIR Workflow with Omniverse
• Building extensions
• Multi-user Collaboration
• Photorealistic rendering with real-time Interaction
PREPARING FOR THE NEXT DECADE OF SCIENTIFIC COMPUTING

INTEGRATING THE SIMULATION +AI AND EXPERIMENT WORKFLOW

ACCELERATING THE SENSOR /EXPERIMENT DATA PROCESSING

BUILDING A DIGITAL TWIN TOWARDS A SCIENCE GRAND CHALLENGE