An Update from ORNL: AI for Science Scalable AI on Summit

David Womble

Oak Ridge National Laboratory

AI won’t replace the scientist, but scientists who use AI will replace those who don’t.*

*Adapted from a Microsoft report, “The Future Computed”

We are at a “tipping point” in AI/ML

Data

- Sensors are ubiquitous
- Data is plentiful
- We are “bit-rich”

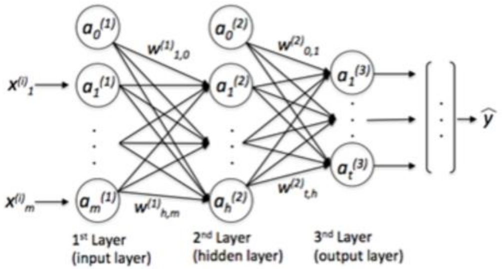
Facilities and data are a distinguishing strength for DOE

Computing

- Computing is “exaflop scale”
- Specialized hardware is being developed for data analytics and “edge” applications

DOE has an HPC mission for science and engineering

Algorithms



- Pre-defined models
- Computationally tractable training for ML

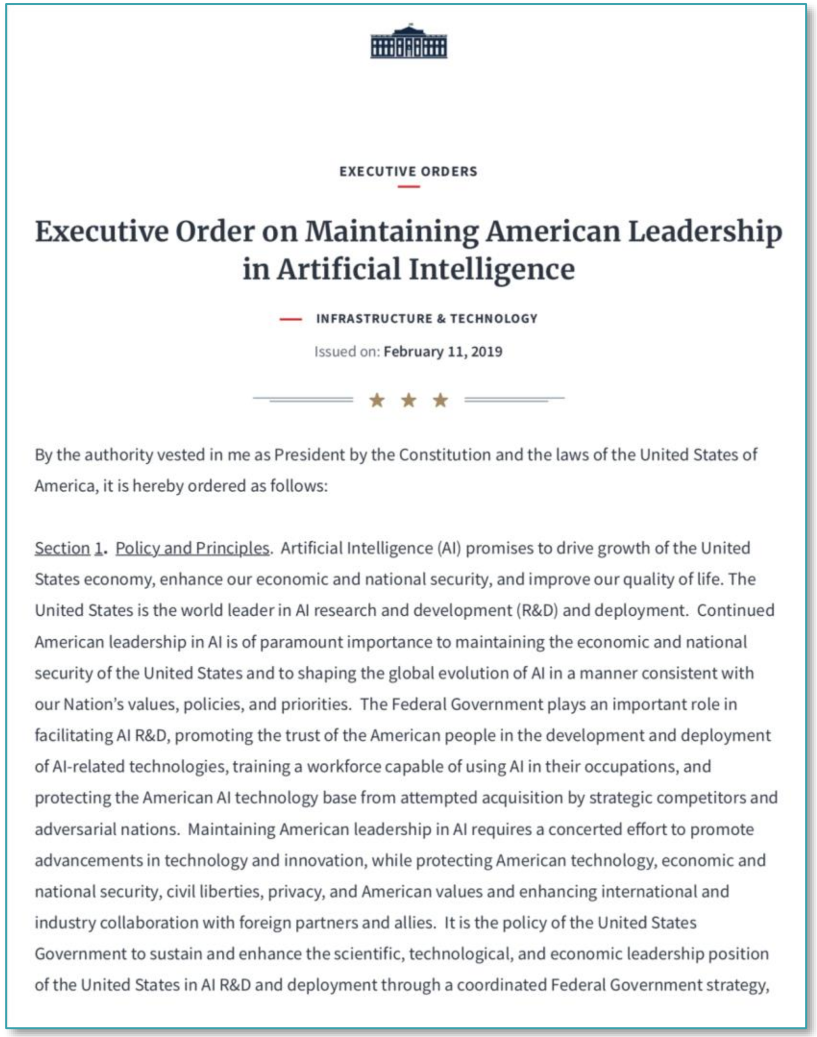
DOE has an HPC mission for science and engineering

Accessibility

- Everyone has a PC and internet access
- A lot of data is open
- Software is open-source

Assurance is “mission critical”

White House Executive Order on AI



EXECUTIVE ORDERS

Executive Order on Maintaining American Leadership in Artificial Intelligence

INFRASTRUCTURE & TECHNOLOGY

Issued on: **February 11, 2019**

★ ★ ★

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

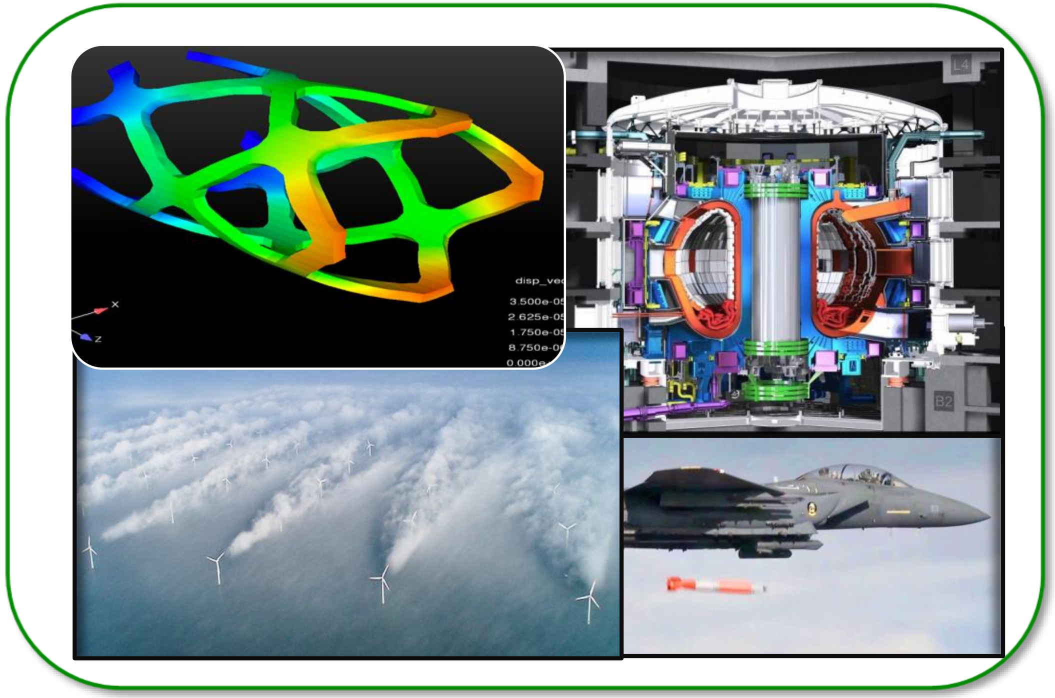
Section 1. Policy and Principles. Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life. The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities. The Federal Government plays an important role in facilitating AI R&D, promoting the trust of the American people in the development and deployment of AI-related technologies, training a workforce capable of using AI in their occupations, and protecting the American AI technology base from attempted acquisition by strategic competitors and adversarial nations. Maintaining American leadership in AI requires a concerted effort to promote advancements in technology and innovation, while protecting American technology, economic and national security, civil liberties, privacy, and American values and enhancing international and industry collaboration with foreign partners and allies. It is the policy of the United States Government to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI R&D and deployment through a coordinated Federal Government strategy,

Policy Statement: Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life.

… leadership requires a concerted effort to promote advancements in technology and innovation, while protecting American technology, economic and national security, civil liberties, privacy, and American values and enhancing international and industry collaboration with foreign partners and allies.

DOE builds on historical missions and touches all areas

- The U.S. AI strategy includes
  1. **Long-term investment in research**
  2. Effective methods for human-AI collaboration
  3. Address ethical, legal and social implications
  4. **Ensure the safety and security of AI Systems**
  5. **Develop shared datasets and environments**
  6. Standards and benchmarks
  7. Understand the AI workforce
  8. Expand public-private partnerships
- DOE will play a key role in AI for science and engineering
  - AI Technology office
  - Research and talent development
  - Data to support science and engineering research



DOE’s Artificial Intelligence and Technology Office

ARTIFICIAL INTELLIGENCE & TECHNOLOGY OFFICE
AITO
UNITED STATES DEPARTMENT OF ENERGY
VIEW ARTICLE

Secretary Perry Stands Up Office for Artificial Intelligence and Technology

This action has been taken as part of the President’s call for a national AI strategy.

SEPTEMBER 6, 2019

Vision:

Transform DOE into a world-leading AI enterprise by accelerating the research, development, delivery, and adoption of AI.

Mission:

The Artificial Intelligence and Technology Office (AITO), the Department of Energy’s center for Artificial Intelligence, will **accelerate the delivery** of AI-enabled capabilities, **scale** the department-wide development and impact of AI, and **synchronize** AI activities to advance the agency’s core missions, **expand partnerships**, and support American AI Leadership.

Office of Science - AI for Science Town Halls

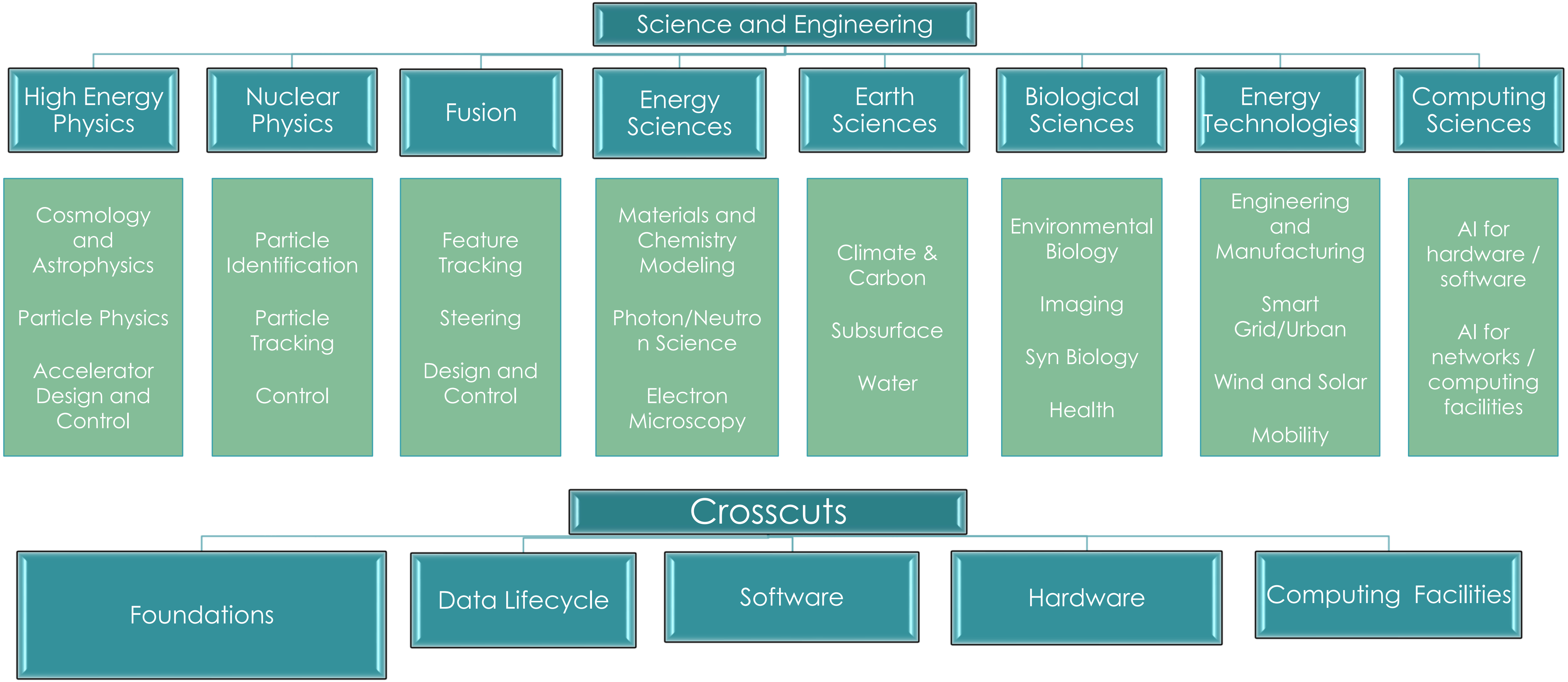
The Integration of modeling and simulation, data analytics and learning

- We are looking for transformational ideas
- What could be the impact of a sustained push on AI in some problem domain?
  - Building superhuman capabilities in science
- What scale
  - Big Problems, Big Pushes, Big Data, Big Systems?
  - Fine grain innovation, many thousands of small teams?
- Coupling to experiments, simulations, user and computing facilities?
- What does “scientific production” look like in this space?

AI for Science Town Halls

- ~1.5 days to capture ideas, problems, requirements and challenges for an AI for Science initiative
- Each townhall
  - 1 plenary, 3 keynotes, half-day breakouts on domains, half-day breakouts on crosscuts
  - All breakouts were consistent, with slight tailoring to accommodate what we learned and local influences
- What problems could be attacked?
- What data, simulations, and experiments do we need?
- What kind of methods, software and math do we need?
- What kind of computer architectures and infrastructure do we need?

Breakouts and Subtopics



Science and Engineering

- High Energy Physics
  - Cosmology and Astrophysics
  - Particle Physics
  - Accelerator Design and Control
- Nuclear Physics
  - Particle Identification
  - Particle Tracking
  - Control
- Fusion
  - Feature Tracking
  - Steering
  - Design and Control
- Energy Sciences
  - Materials and Chemistry Modeling
  - Photon/Neutron Science
  - Electron Microscopy
- Earth Sciences
  - Climate & Carbon
  - Subsurface
  - Water
- Biological Sciences
  - Environmental Biology
  - Imaging
  - Syn Biology
  - Health
- Energy Technologies
  - Engineering and Manufacturing
  - Smart Grid/Urban
  - Wind and Solar
  - Mobility
- Computing Sciences
  - AI for hardware / software
  - AI for networks / computing facilities

Crosscuts

- Foundations
- Data Lifecycle
- Software
- Hardware
- Computing Facilities

Science Breakthroughs

What: is the challenge problem?

Why: is this important to science, society, etc.?

How: what kind of AI is critical and why?

Scale: what is the data size/rate, compute cost, etc.?

Timeframe: is this a 3,5, or 10-year goal?

Crosscut Challenge Highlight

What problem are you solving?

Why is this important?

Which applications need it?

Why DOE? How does this fit into DOE expertise / facilities / team science?

The DOE and Office of Science in FY20

- Expect a strategic plan from AITO
- Office of Science has $71M guidance in FY20 budget
  - Most office will release calls
  - ASCR guidance is $36M between math, computer science and partnerships
- There will also be calls from many other programs
  - Some calls will be “inverted”

AI at ORNL

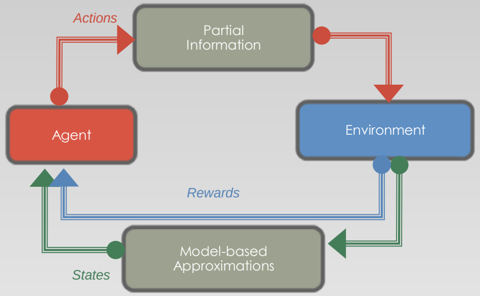
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

ORNL Strategic Directions in AI/ML

Data

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

Learning

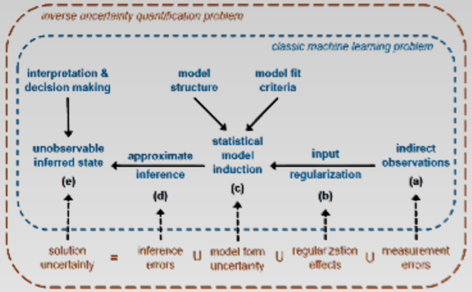


- Physics informed
- Reinforcement learning
- Adversarial networks
- Representation learning and multi-modal data
- "Foundational math" of learning

Scalability

- Algorithms, complexity and convergence
- Levels of parallelization
- Mixed precision arithmetic
- Communication
- Implementations on accelerated-node hardware

Assurance



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

Workflow

- Computing at the edge
- Compression
- Online learning
- Federated learning
- Infrastructure
- Augmented intelligence
- Human-computer interface

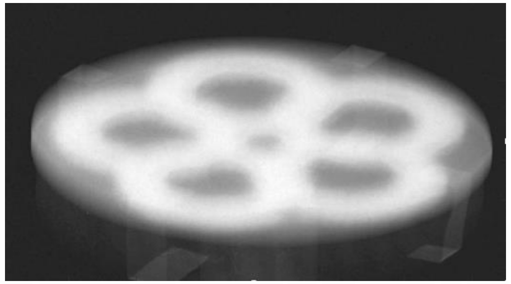
What is the AI Initiative at ORNL?

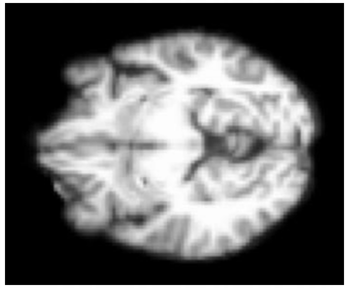
FY21: AIRES (AI for Robust Engineering and Science)

AI at the edge (?)

Image Analytics

Dramatically improve performance of instruments and systems for image analysis through probabilistic modeling on data-derived image manifolds.



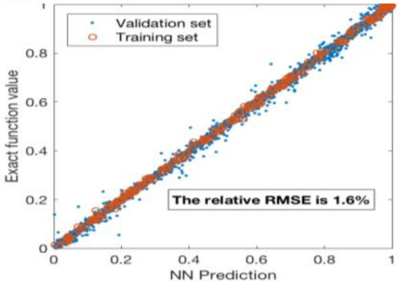


Scalability

Develop scalable machine learning algorithms and tools that use exascale computing capabilities effectively.

ML Foundations

Exploit ORNL's math expertise to develop fundamental ML theories for scientific applications, work the ORNL's HPC experts to transfer the new ML theories to scalable ML capabilities, and demonstrate the advancement of the new capabilities in ORNL's core domain areas.



OAK RIDGE National Laboratory

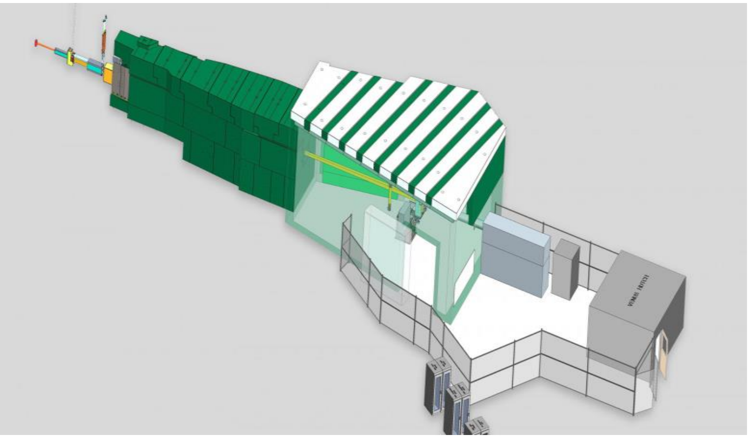
HPC for Scalable Image Analytics

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

U.S. DEPARTMENT OF ENERGY

Summit-scale U–Net Training

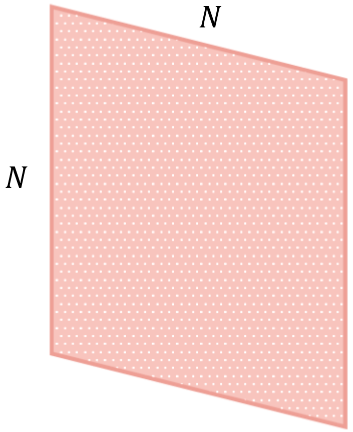
- Goal is not just faster, but also better.
- Satellite images collected at high-resolutions (30-50 cm) yield very large 10,000 x 10,000 images.
- Training ML models on these large high-resolution images is extremely challenging.
- Accurate ML models are needed to resolve multi-scale objects (buildings, solar panels, land cover details).
- U-Net models preferred -- good for training with limited labeled data.
- At present, requires many days to train a single model (even on special-purpose DL platforms like DGX boxes).
- Hyperparameter tuning of these models take much longer.



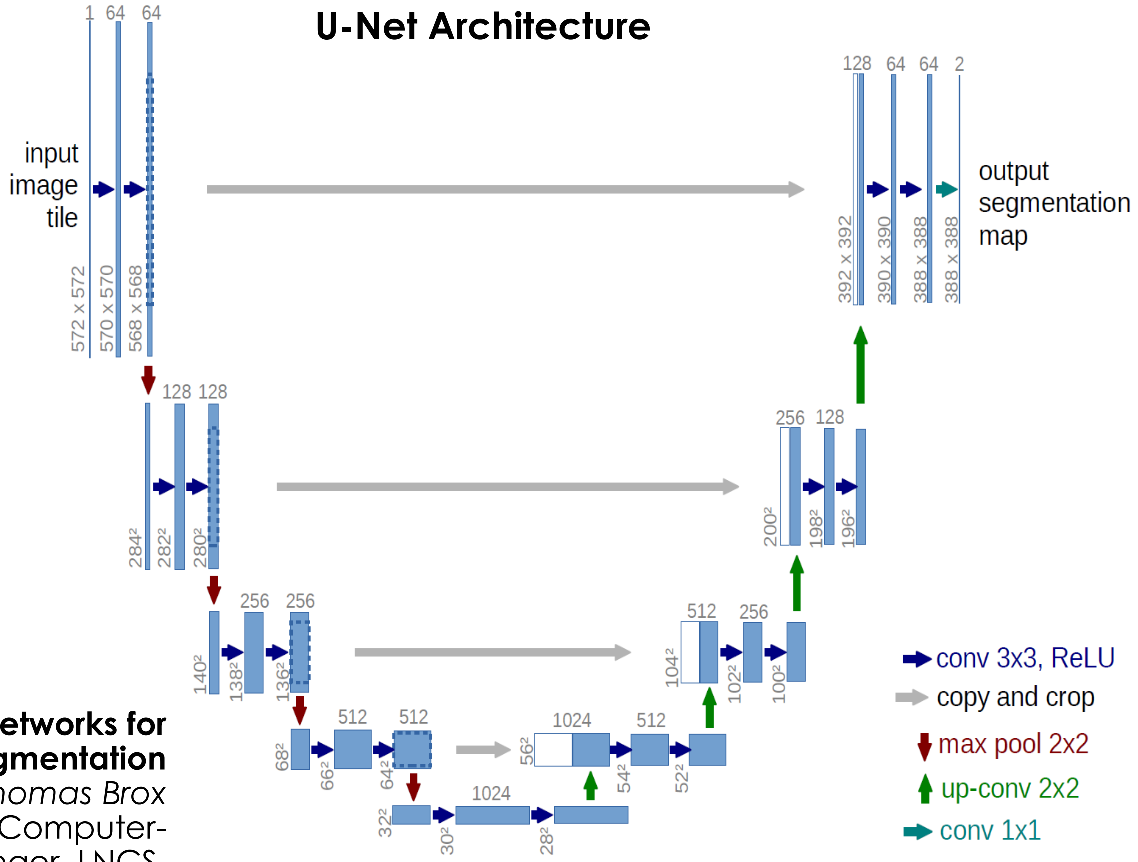
ML technologies developed will be here applicable to other large-scale image analytics domains (e.g., as anticipated from the VENUS neutron imaging instrument).

Semantic Segmentation with U-Net

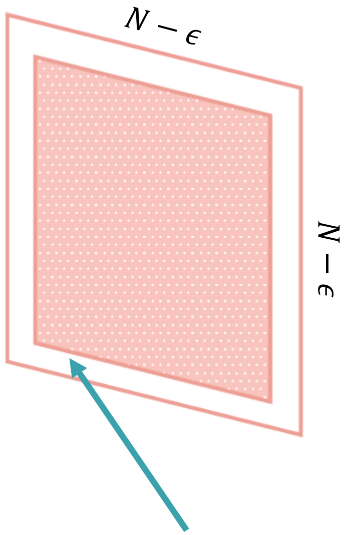
Input Image



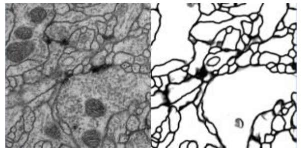
U-Net Architecture



Output Image



- Refer to this as the halo region.
- Halo width is a function of U-Net architecture (depth, channel width, filter sizes, etc.).
- Halo width determines the receptive field of the model.
- Larger the receptive field, wider the length-scales of identifiable objects.

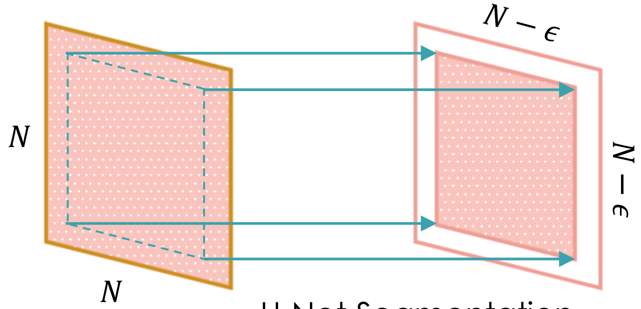


U-Net: Convolutional Networks for Biomedical Image Segmentation
Olaf Ronneberger, Philipp Fischer, Thomas Brox
Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241, 2015.

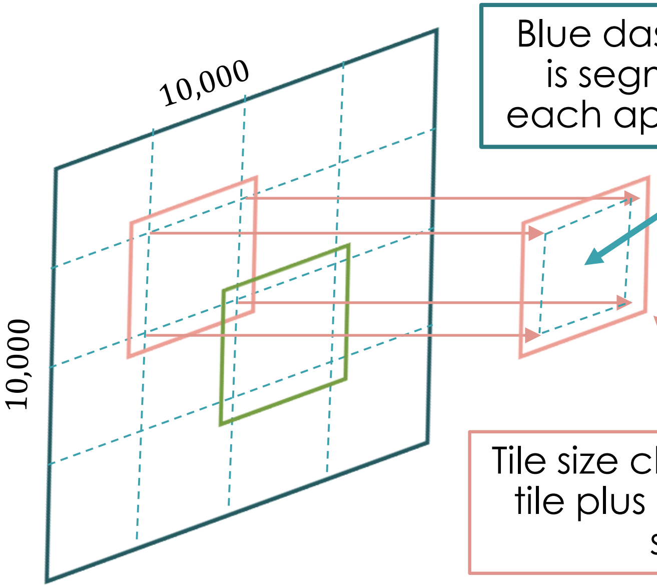
Scalable Sample Parallel Segmentation

Leveraging Summit's Vast GPU Farm

- Given a $N \times N$ image, U-Net segments a $(N-\epsilon) \times (N-\epsilon)$ inset square.



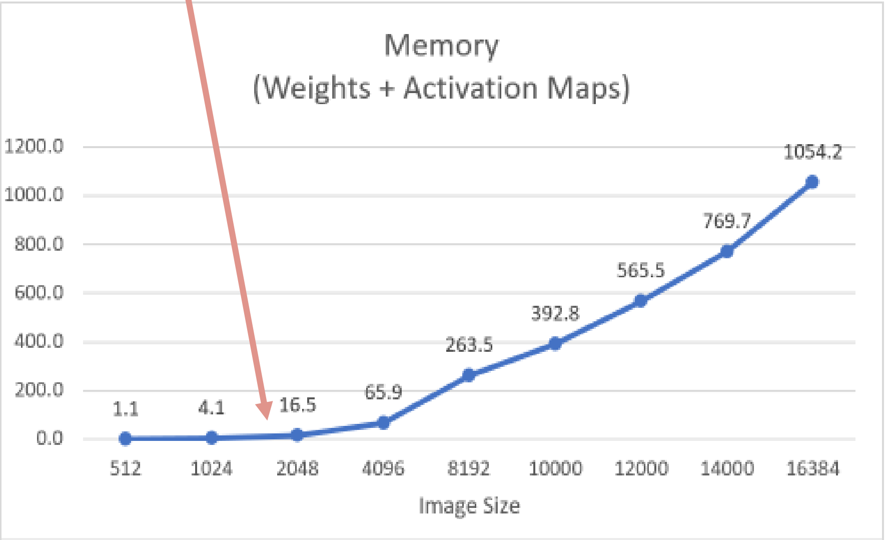
U-Net Segmentation



Blue dashed square is segmented for each appended tile.

Tile size chosen such that appended tile plus model parameters fit on a single Summit GPU.

- Partition each 10,000 x 10,000 image sample into non-overlapping tiles.
- Append an extra halo region of width ϵ along each side of each tile.
- Assign each appended tile to a Summit GPU. Use U-Net to segment appended tile.
- Each GPU segments an area equal to that of the original non-overlapping tile.



+100X Faster U-Net Training

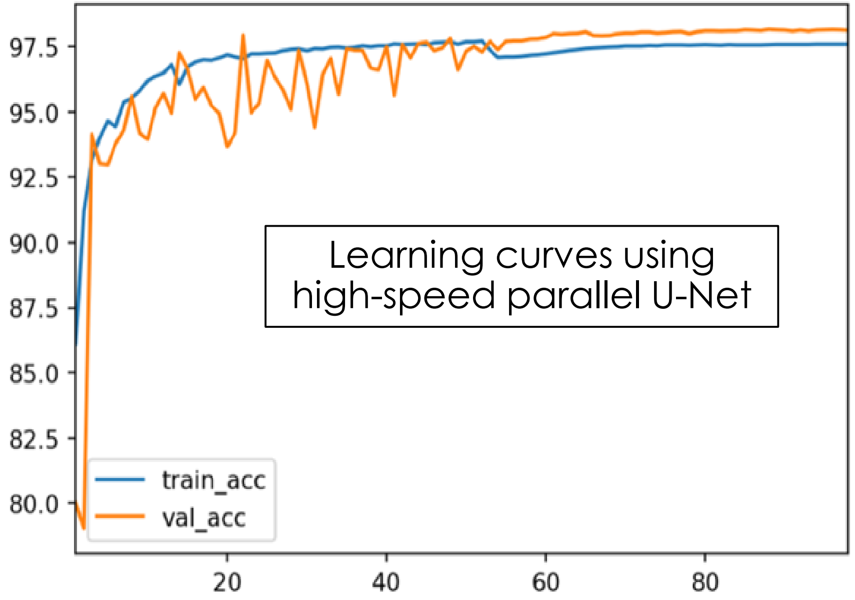
- Optimal tiling for each 10,000 x 10,000 sample image was found to be 8 x 8.

- Each 1250 x 1250 tile was appended with a halo of width 92 and assigned to a single Summit GPU.
    - ➢ 10 Summit nodes to train each 10,000 x 10,000 image sample.

- A U-Net model was trained on a data set of 400 very large (10,000 x 10,000 x 4) satellite images, collected at 30-50 cm resolution.

- The training time per epoch was shown to be ~**12 seconds** using **1200 Summit GPUs** compared to ~**1,740 seconds** on a **DGX-1**.

- Initial testing revealed no appreciable loss of training/validation accuracy using the new parallel framework.

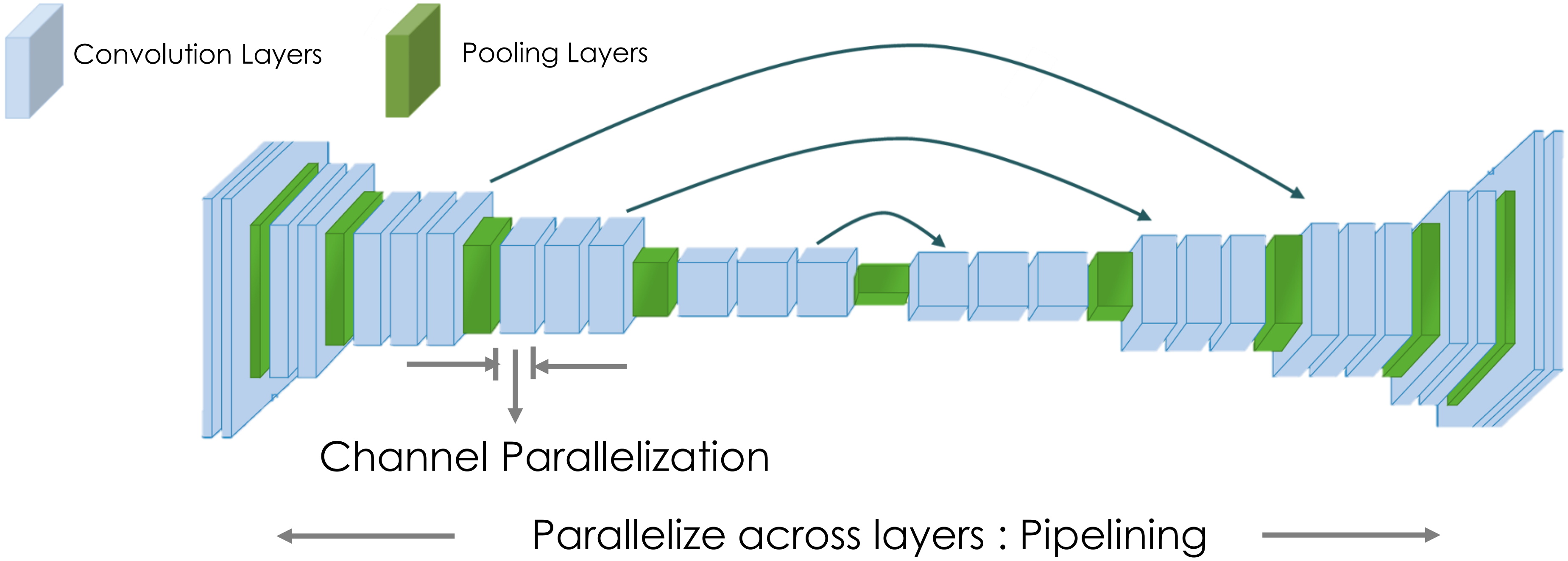


Learning curves using high-speed parallel U-Net

Image credit: Maxar Technologies

Limitations

- Does not improve the receptive field of the original U-Net architecture.
  - Limits the ability to identify objects of widely varying dimensions (say, people from buildings).
- Larger receptive fields require deeper U-Net architectures.
- Necessitates larger models that cannot be fit on a single GPU.



Channel Parallelization

Parallelize across layers : Pipelining

Stable Code Development Environment on Summit

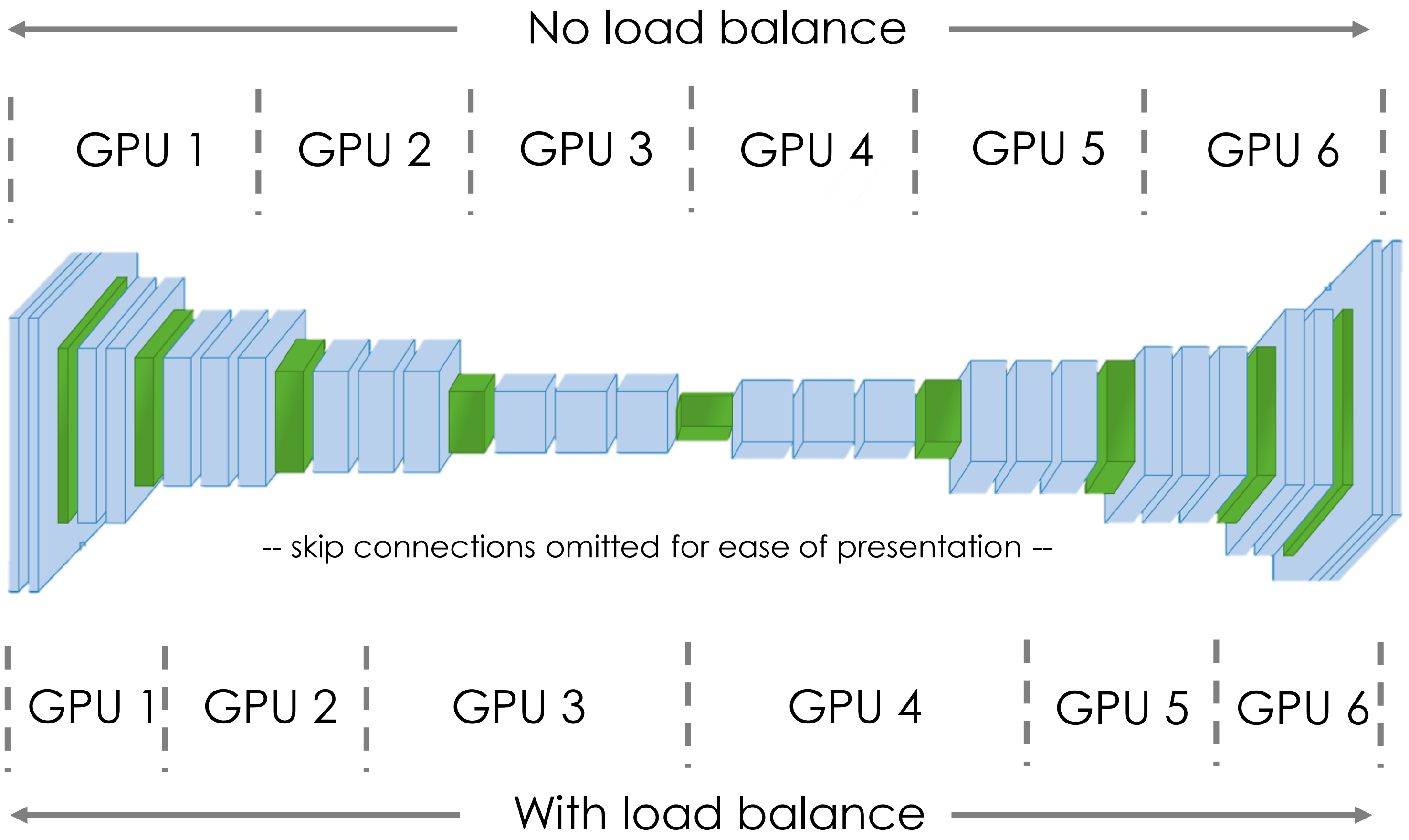
- Issue:
  - Need sustained Summit-support and PowerPC/V100-optimized libraries.
  - Avoid/reduce package dependencies.
  - A programming environment with flexibility to ”play around” with at Summit-scale.

- DiXN Approach:
  - IBM’s Watson Machine Learning (WML), Community Edition (1.6.2-0).
  - IBM WML includes standard deep learning packages (TensorFlow/PyTorch), which are also optimized for Summit.
  - IBM WML CE supports IBM’s distributed deep learning (DDL) package, which is implemented on top of MPI and optimized for Summit architecture, including network topology.
  - Horovod+MPI4Py for communication management.

IBM Watson™

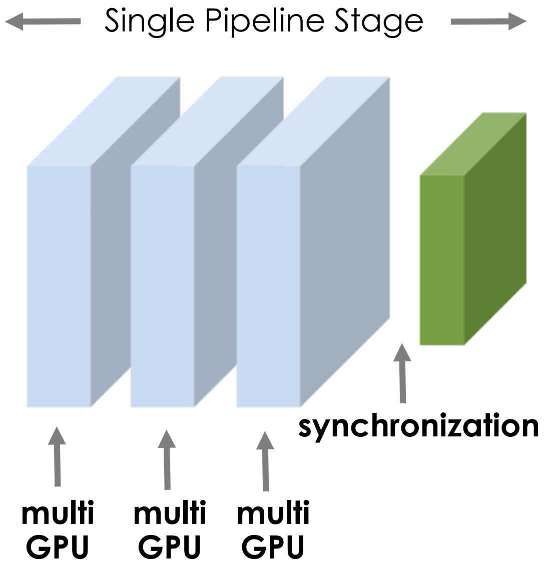
Pipelined DiXN

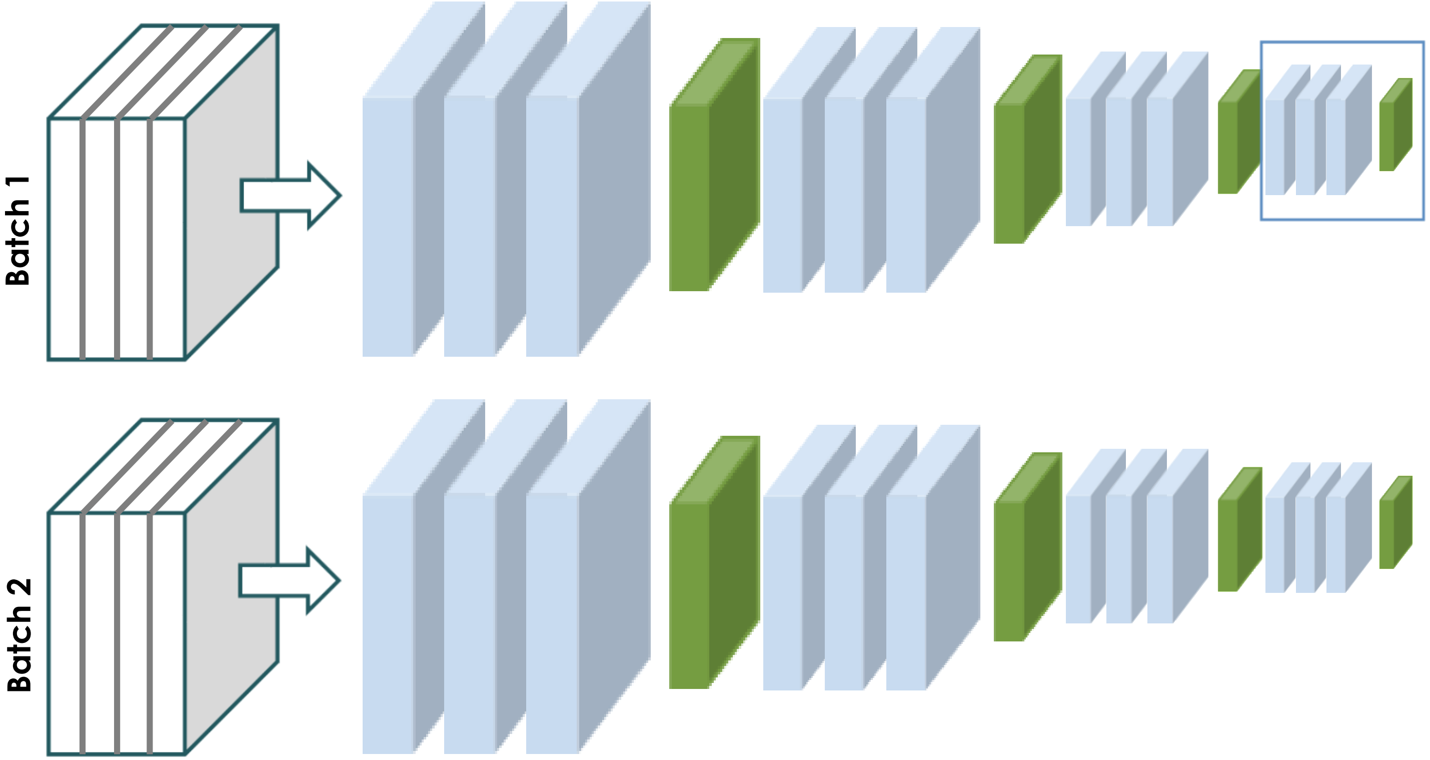
- Story 2 (intranode) will be delivered in two releases.
- The first release will partition layers equally amongst the GPUs available within a node (6 for Summit).
- This partitioning will be load imbalanced.
- Primary focus on building communication infrastructure for pipelined execution of an X-Net.
- The second release will integrate load balancing algorithms.



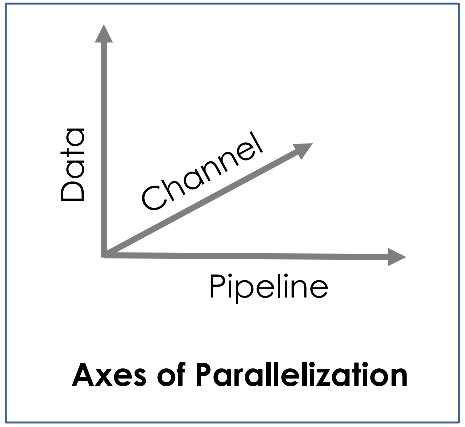
Channel Parallel DiXN

- Each convolutional layer exhibits fine-grained parallelism suitable for SIMD parallelism.
- Partition convolution computations per layer across multiple GPUs.
- Synchronizations required before max-pooling layer.





Data + Model Parallel Execution at Summit-scale



Axes of Parallelization

2020
CALL FOR PAPERS

The call for abstracts is open now through March 27!

Join us in Kingsport, Tennessee from August 25 to 27 for the Smoky Mountains Computational Sciences and Engineering Conference (SMC2020), a premier event focused on computational science and engineering.

SESSION TOPICS:

- Computational Applications
- System Software
- Experimental/Observational Applications
- Deploying Computation
- Scientific Data Challenges

Learn more at
http://smc.ornl.gov

Thank you