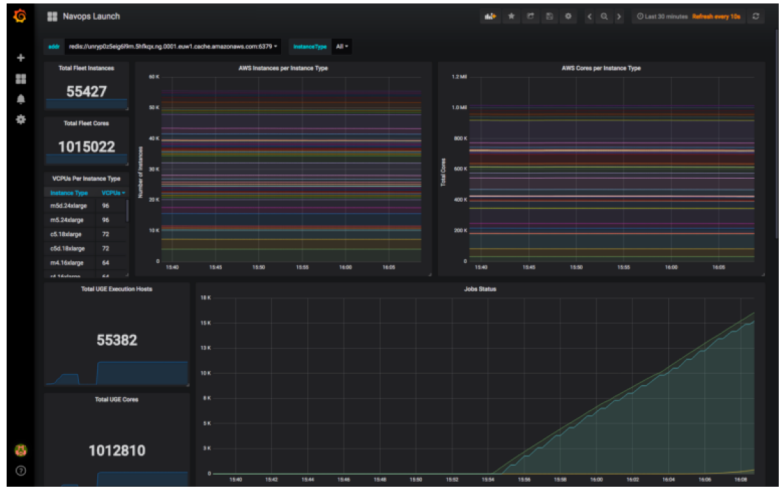
https://www.nextplatform.com/2018/08/29/cascade-lake-heart-of-2019-tacc-supercomputer/

LARGE-SCALE DEEP LEARNING AS A BENCHMARK

Pradeep K Dubey

Intel Fellow and Director Parallel Computing Lab

Intel Labs

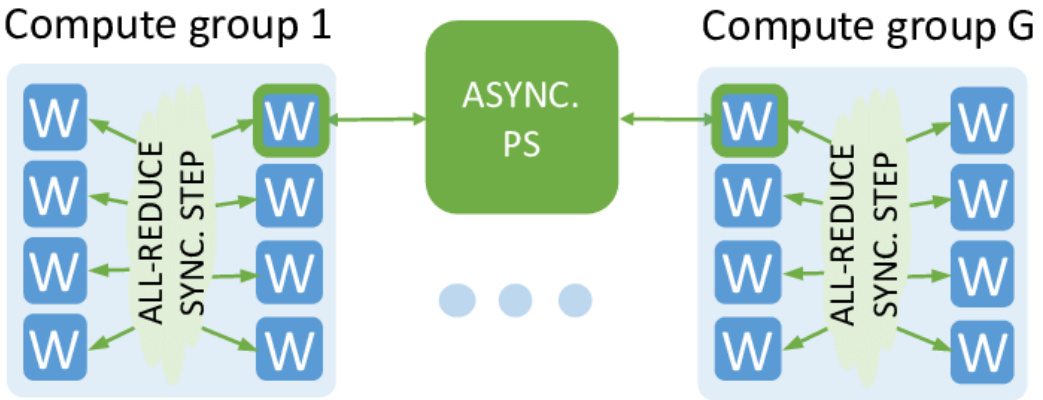


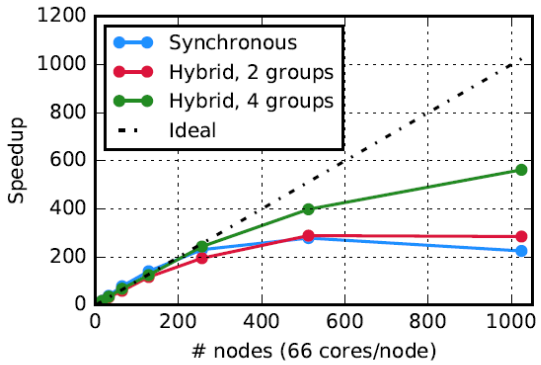
https://insidehpc.com/2018/06/univa-deploys-million-core-grid-engine-cluster-aws/

OUR EXPERIENCE SCALING DEEP LEARNING: DEEP LEARNING AT 15PF *

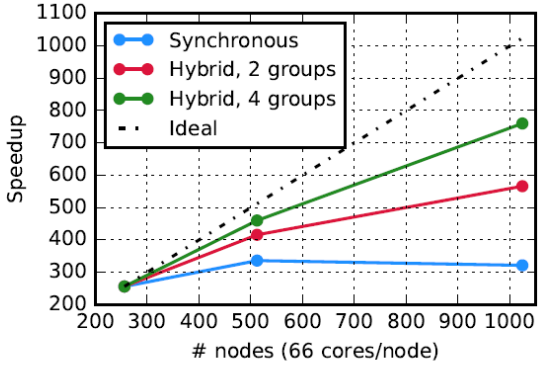
- Joint work between NERSC, Stanford University and Intel on Cori Phase II
- Novel approach to distributed SGD: synchronous within the group, asynchronous across the groups
- Record scaling: in terms of number of nodes collaboratively training the same model (9600 KNL)
- Record peak performance: ~15PF
- Communication approach: MLSL for intragroup communication, MPI for intergroup
- The mechanism is available in IntelCaffe/MLSL
- Real-world pattern classification problems in HEP and Climate Science

http://www.nersc.gov/users/computational-systems/cori/configuration/





(a) HEP



(b) Climate

* https://dl.acm.org/citation.cfm?id=3126916

WHAT SHOULD WE NEED TO CONSIDER IN A LARGE-SCALE DL BENCHMARK

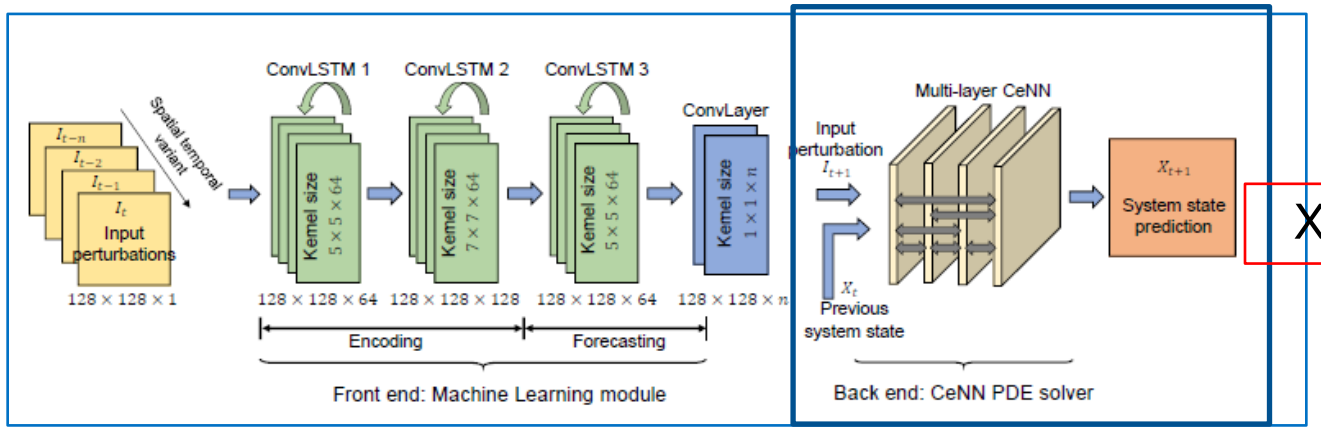
- Clearly state the goal: *Ranking the compute infrastructure goodness*
- Scale matters – 1000s of nodes
- End-to-end time-to-train to a given level of accuracy
- Consider TCO as well
- Include public cloud – now offering millions of cores

- Proxy of a real-world, forward-looking application
  - Not another ImageNet ☺
- Challenging enough DL training with spelt-out network and dataset
- Allow flexibility of data-model parallelism, inter-node communication, precision, etc.
  - All such need to be disclosed
- Open source code, reproducible results

TREND 1A:

NEURAL NETWORKS GETTING AUGMENTED

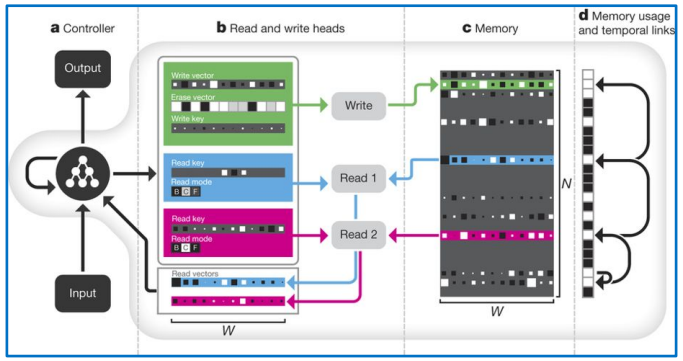
NN + X



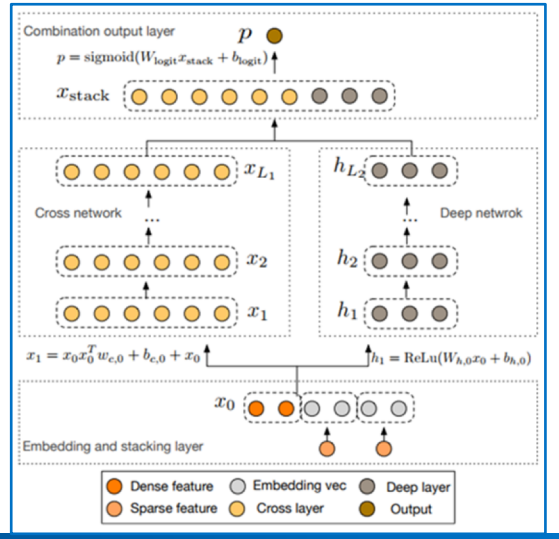
X = Coupled PDEs

Hybrid learning of with neural networks and coupled dynamic system (PDEs) for heat dissipation and fluid dynamics [1]

NN + Memory



Memory needed to perform Turing-complete operations. DeepMind work on differentiable memory [2]



Large embedding tables mapping sparse feature vector to dense vectors – Requires several TBs of memory [3]

TREND 2:

FOCUS SHIFTS FROM PATTERNS TO ANOMALIES

CYBER-SECURITY
SOCIAL MEDIA

REAL-TIME MONITORING
INTERNET OF THINGS (IOT)

AUTONOMOUS DRIVING
DATA ANALYTICS

DATA CENTERS
WEATHER / ASTRONOMY

Anomaly Detection Is *Everywhere*

NEW EVALUATION MODEL

Expressive, Flexible, Extensible

- **Superset of:**
  - Classical model
  - Other state-of-the-art evaluators (NAB)
- ***NIPS '18 Spotlight***
- **Other NIPS'18 Highlight:**
  - Intel for the first time in the league of >1% of accepted paper affiliations

Precision and Recall for Time Series

Nesime Tatbul
Intel Labs and MIT
tatbul@csail.mit.edu

Tae Jun Lee
Microsoft
tae_jun_lee@alumni.brown.edu

Stan Zdonik
Brown University
sbz@cs.brown.edu

Mejbah Alam
Intel Labs
mejbah.alam@intel.com

Justin Gottschlich
Intel Labs
justin.gottschlich@intel.com

Abstract

Classical anomaly detection is principally concerned with *point-based anomalies*, those anomalies that occur at a single point in time. Yet, many real-world anomalies are *range-based*, meaning they occur over a period of time. In this paper, we present a new model that more accurately measures the correctness of anomaly detection systems for range-based anomalies, while subsuming the classical model's ability to classify point-based anomaly detection systems.

1 Introduction

Anomaly detection (AD) is the process of identifying non-conforming items, events, or behaviors. The proper identification of anomalies can be critical for many domains. Some examples are early diagnosis of illness and disease [18], threat detection for cyber-attacks [3], or safety analysis for self-driving cars [29]. Many real-world anomalies can be detected in time series data. Therefore, systems that detect anomalies should reason about them as they occur over a period of time. We call such events *range-based anomalies*, which are a subset of both contextual and collective anomalies [9]. More precisely, a *range-based anomaly* is an anomaly that occurs over a consecutive sequence of time points, where no non-anomalous data points exist between the beginning and the end of the anomaly. The standard metrics for evaluating anomaly detection algorithms today, *Recall* and *Precision*, have been around since the 1950s, originally formulated to evaluate document retrieval algorithms by counting the number of documents that were correctly returned against those that were not [7].

Formally defined as follows, Recall and Precision are a good match for single-point AD [1] (where TP, FP, FN are the number of true positives, false positives, false negatives, respectively):

$$Recall = TP \div (TP + FN) \quad (1)$$

$$Precision = TP \div (TP + FP) \quad (2)$$

Informally, *Recall* is the rate at which a system can identify anomalies without mispredicting any anomalous events. *Precision* is the rate a system can identify anomalies without mispredicting non-anomalous events. In this sense, *Recall* and *Precision* are complementary. This characterization proves useful when they are combined, such as in the $F_1 score$, which is their harmonic mean. Such combinations help gauge the quality of both anomalous and non-anomalous predictions. While useful for point-based anomalies, classical recall and precision suffer from the inability to represent domain-specific time series anomalies. This has a negative side-effect on the advancement of AD systems. In particular, many time series AD systems' accuracy is being misrepresented, because point-based recall and precision are being used to measure their effectiveness for range-based anomalies. Moreover, the need to accurately identify time series anomalies is growing in importance due to the explosion of streaming and real-time systems [2, 6, 14, 23, 28, 31]. To address this, we redefine recall and precision to encompass range-based anomalies. Unlike prior work [2, 21], our mathematical

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

NIPS '18 (Spotlight): https://arxiv.org/abs/1803.03639 (non-camera-ready version)

TREND 3: MOVING BEYOND PERCEPTION

FOCUS SHIFTS FROM SPARSE AND GRAPH ANALYTICS

Many Big Data sets can be represented as a graph: Social networks, IP network traffic, road networks, physics models, etc.

Graph analytics can reveal interesting information:
detecting patterns and clusters, shortest path calculation, search problems

Differences with ‘classic’ HPC:

- Sparse data: connection matrix has small fraction of non-zeros
- Light computations: walking through graph makes most algorithms memory bound
- Data dependent: process time depends on number of neighbors, which can be highly unbalanced