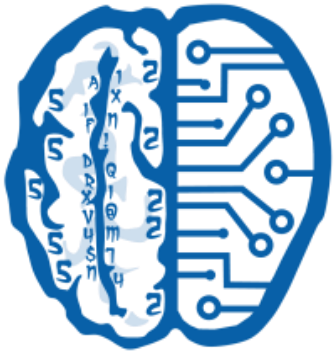


ICRI-CI

Intel Collaborative Research Institute
Computational Intelligence

intel

Capstone Overview

Architecture for Big Data & Machine Learning

Debbie Marr

ICRI-CI 2016 Retreat, May 24, 2016

INTEL LABS

Deliver breakthrough innovations to fuel Intel's growth and technology leadership

Accelerators

Memory Traffic Reduction

Memory Intensive Arch.

Context-based Prefetching

Deep Learning

SimNets

Distributed Methods for Deep Learning

Scene Understanding

Saliency Estimation

Statistics of Depth Images

Arguments for Persuasive Discussion

Universal Semantics

Transcript Quality

Inference for NLP

Relations and Events

Extraction Knowledge Graphs

Hybrid Models

Syntactic & Semantic Reranking

Language Modeling

2nd-order Embedding

Mental Phenotyping

Reinforcement Learning

ICRI-CI

Intel Collaborative Research Institute Computational Intelligence

Distributed Deep Learning Library

Conversational Speech Understanding

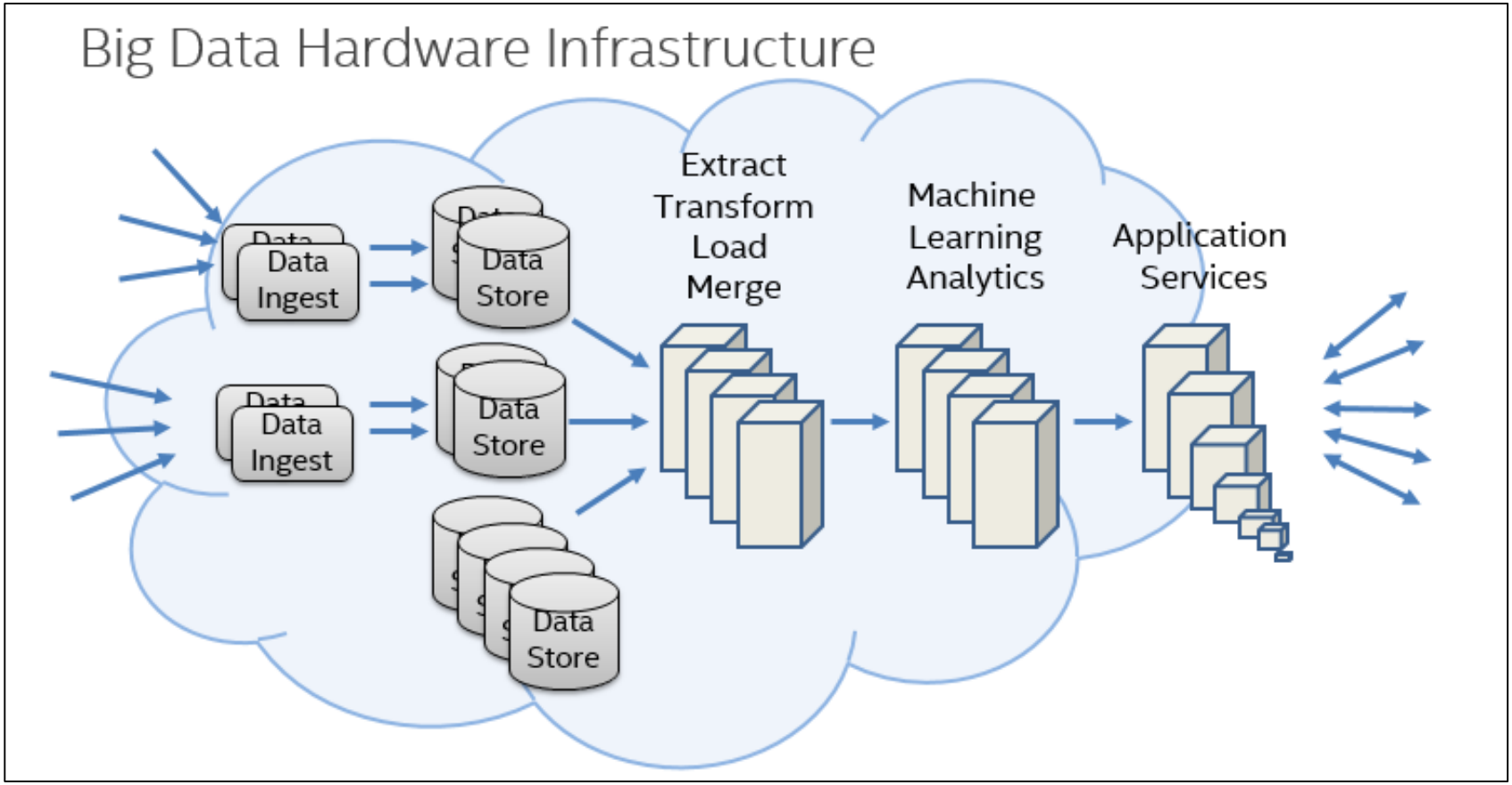
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence

Accelerators for Big Data and Machine Learning

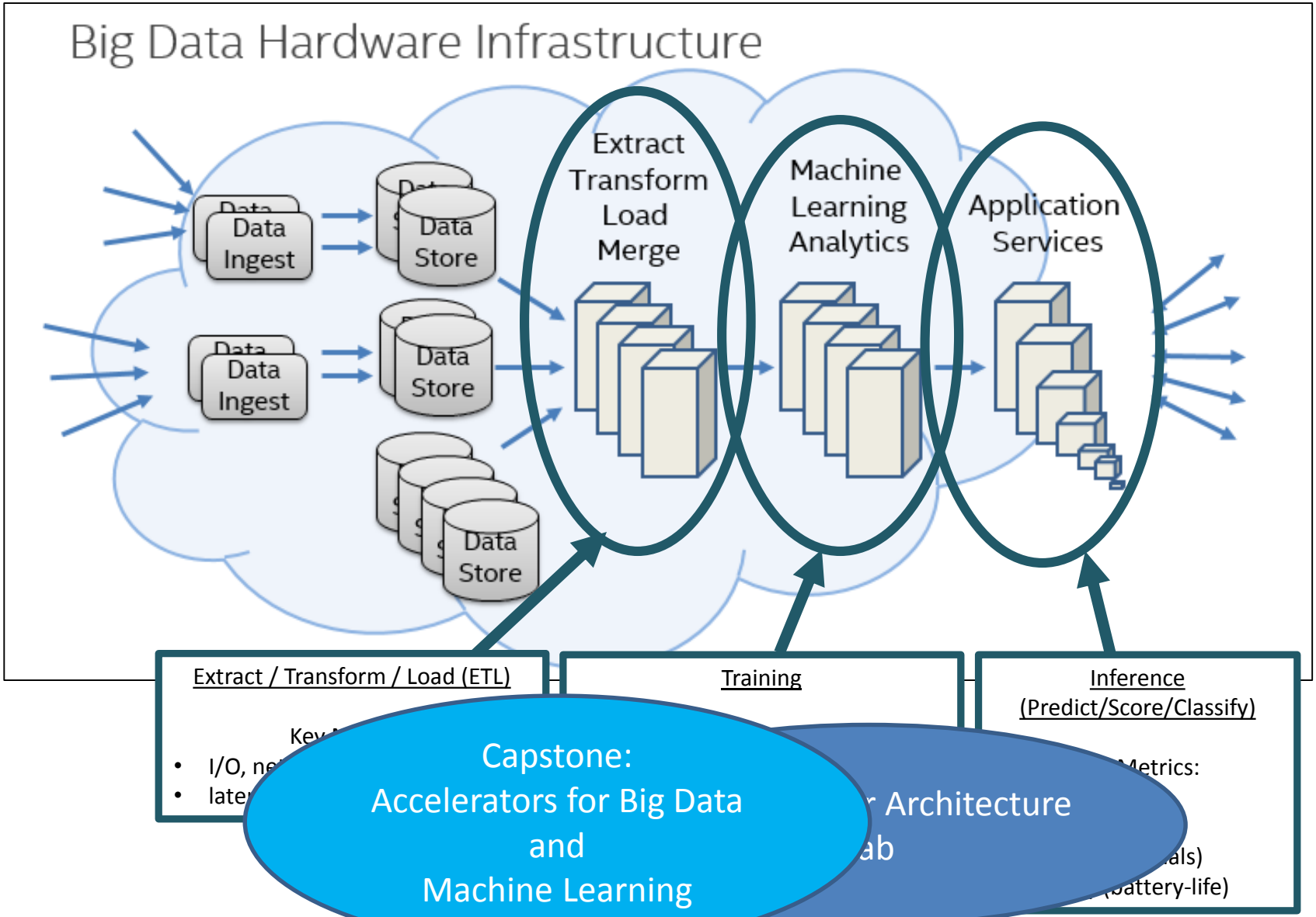
INTEL LABS

Deliver breakthrough innovations to fuel Intel's growth and technology leadership

Big Data / Machine Learning Hardware Infrastructure View

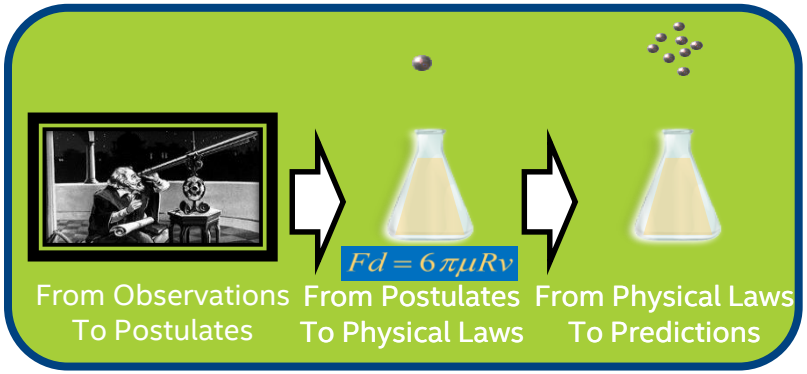


AAL & ICRI-CI Accelerator Investments

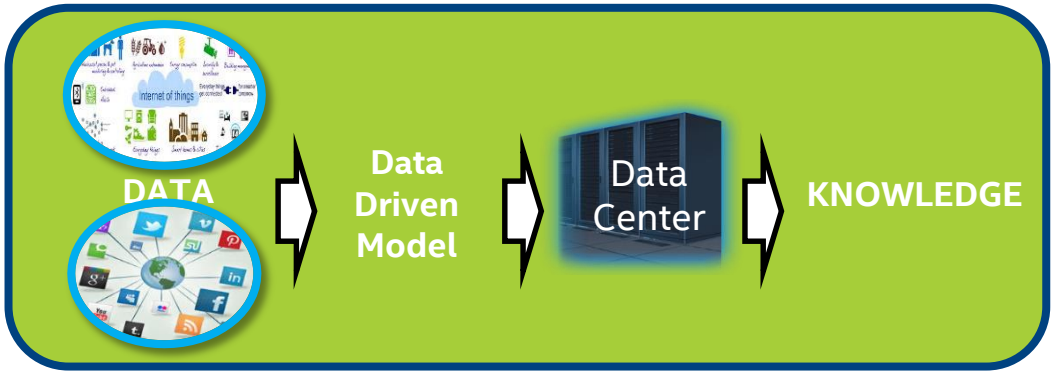


Humans vs. Compute Systems

Human Knowledge Acquisition



Compute Systems Knowledge Acquisition



Brain:
Compute efficiency: 4-5 Gop/W
Power of compute vs. communication: 50:50

Computing systems:
Compute efficiency: 4-5 Gop/W achievable
Power of compute vs. communication: 10:90

Refactor compute systems around data & communication

- This year data center flops: ~1 Exaflop (10^18)
- This year internet data rate: ~1.6 Zetabyte/s (10^14)
  - ~10^4 flops to process 1 byte of internet traffic
- Compute doubling every 1.5 years
- Data storage doubling every 1 year
- Data analytics fastest growing class of data center workload

Source: Pradeep Dubey

Computing Systems: Don't work harder, work smarter

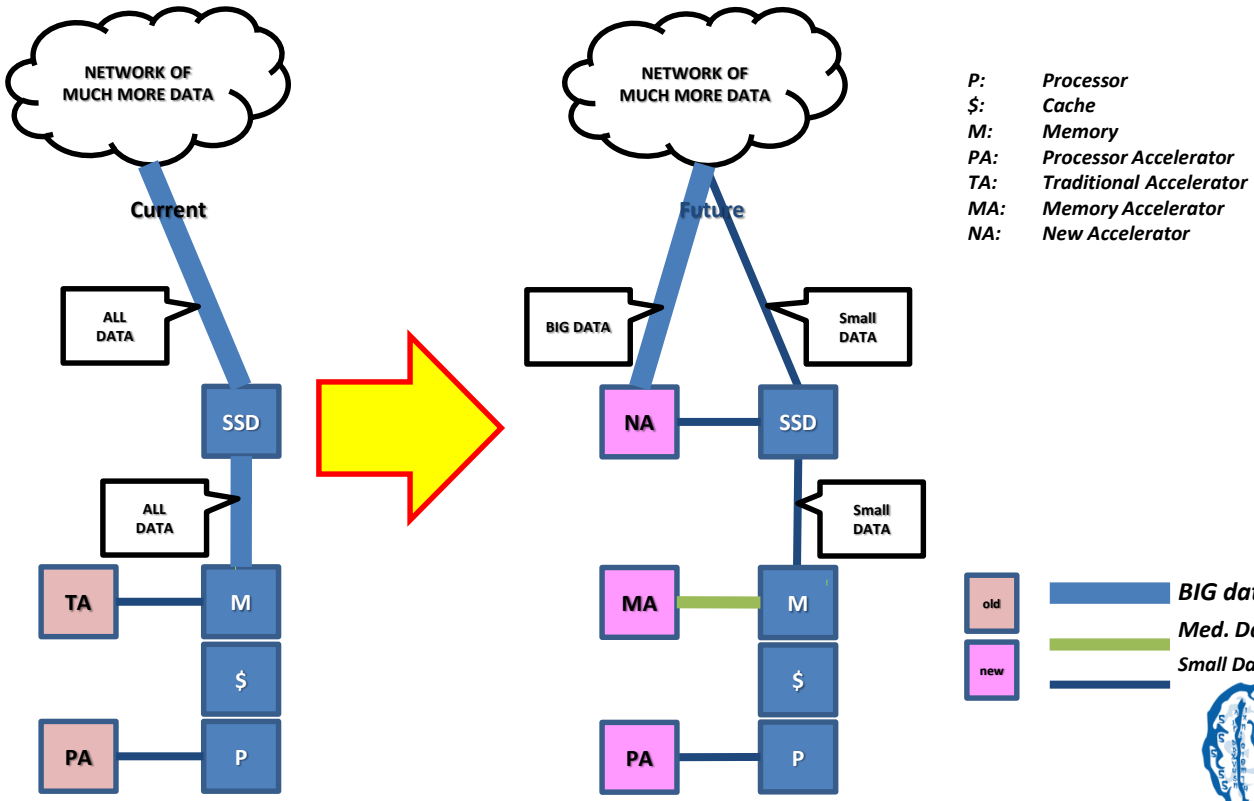
- Working harder:
  - Peak FLOPs
  - Peak Bandwidth
- Working smarter:
  - Re-think compute, storage, bandwidth
    - Move data to compute, or compute to data?
    - Dynamically adapt to changing needs for data, compute, bandwidth
    - Leverage the sparsity in the models to our advantage.

Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. Data movement in/across nodes
2. Computation placed in the storage & network hierarchy
3. New accelerators for big data
4. Applications and usage of new memory technologies (e.g. memristors)
5. Leveraging ML algorithms for new microarchitectures



ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence

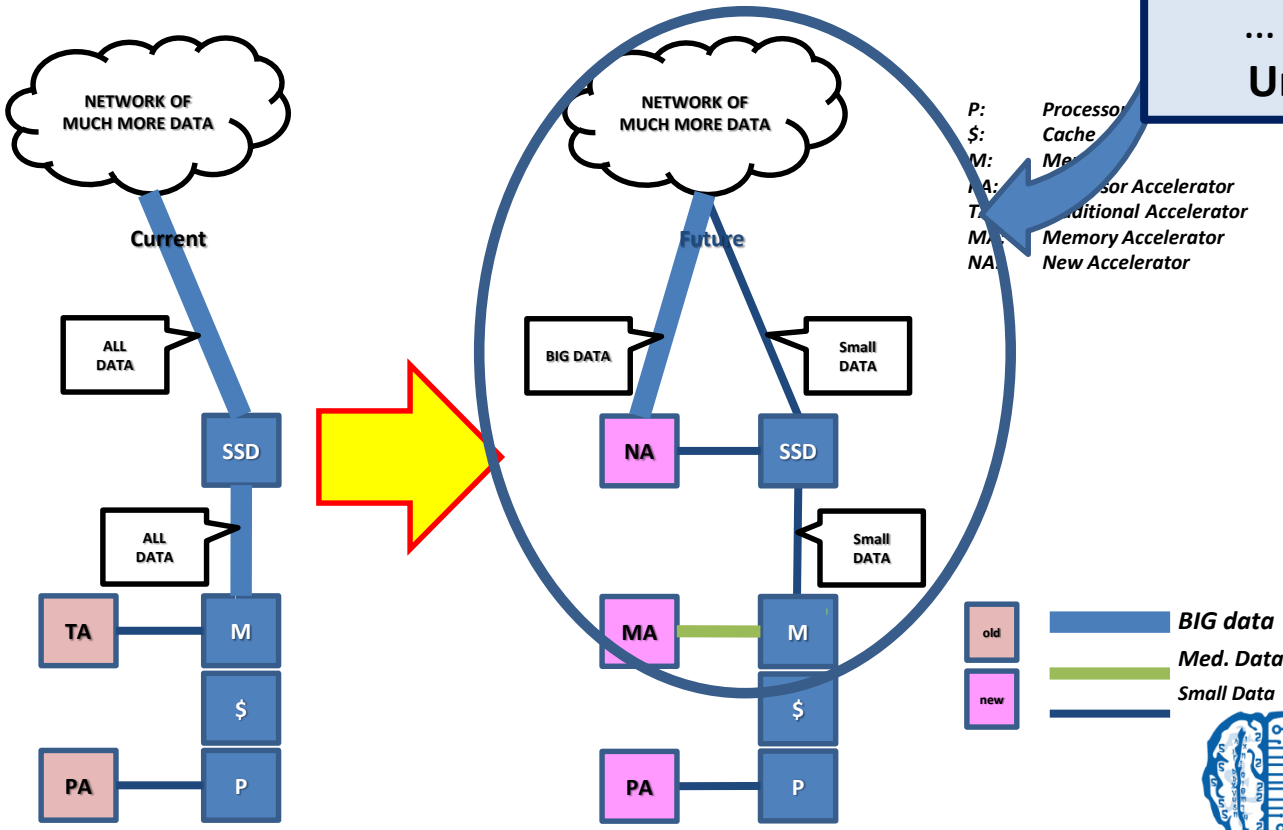
Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. **Data movement in/across nodes**
2. **Computation placed in the storage & network hierarchy**
3. New accelerators for big data
4. Applications and usage of new memory technologies (e.g. memristors)
5. Leveraging ML algorithms for new microarchitectures

"Process-in-storage … or not?"
Uri Weiser



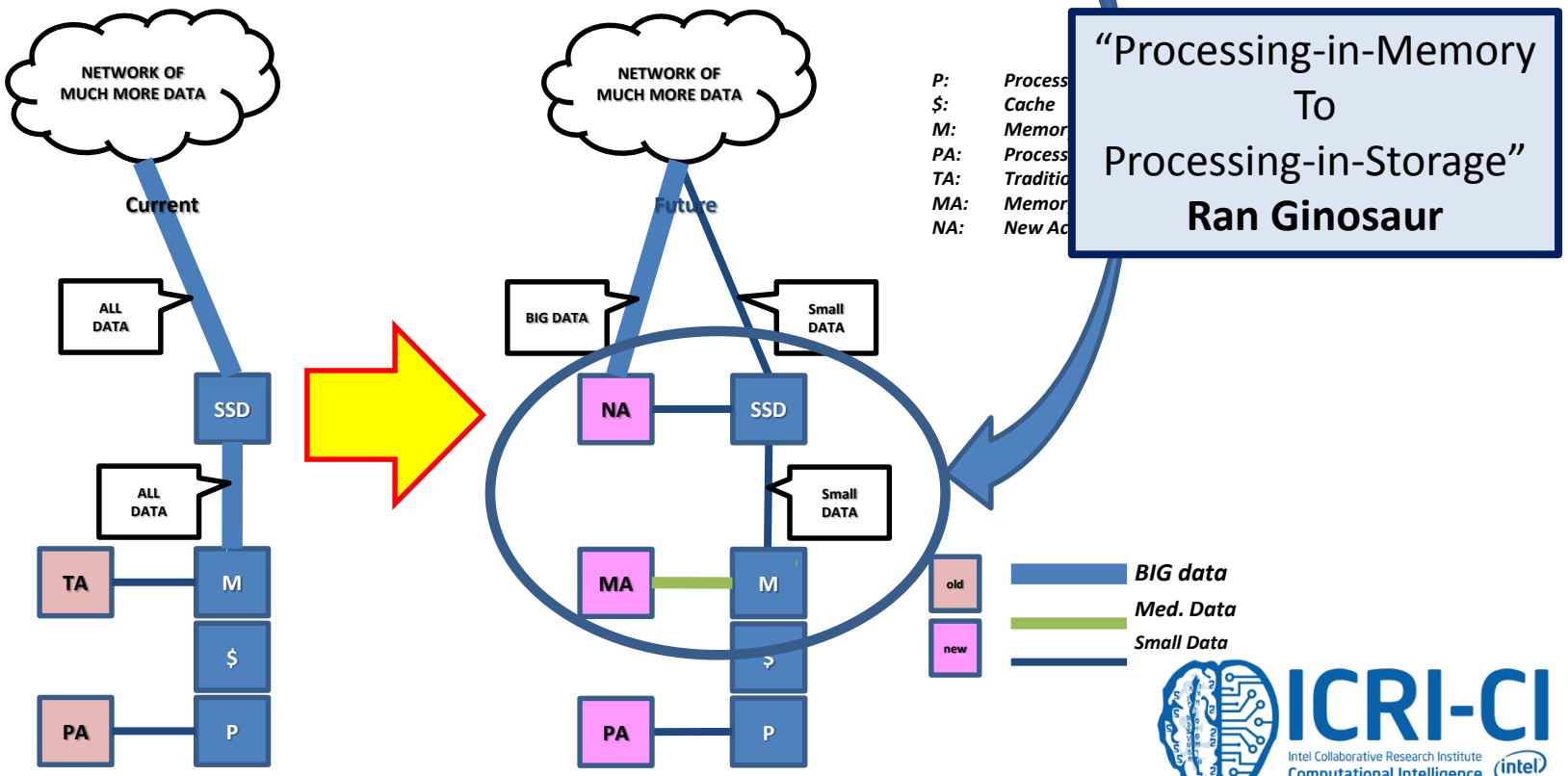
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. Data movement in/across nodes
2. **Computation placed in the storage & network hierarchy**
3. **New accelerators for big data**
4. Applications and usage of new memory technologies (e.g. memristors)
5. Leveraging ML algorithms for new microarchitectures



"Processing-in-Memory To Processing-in-Storage"
Ran Ginosaur

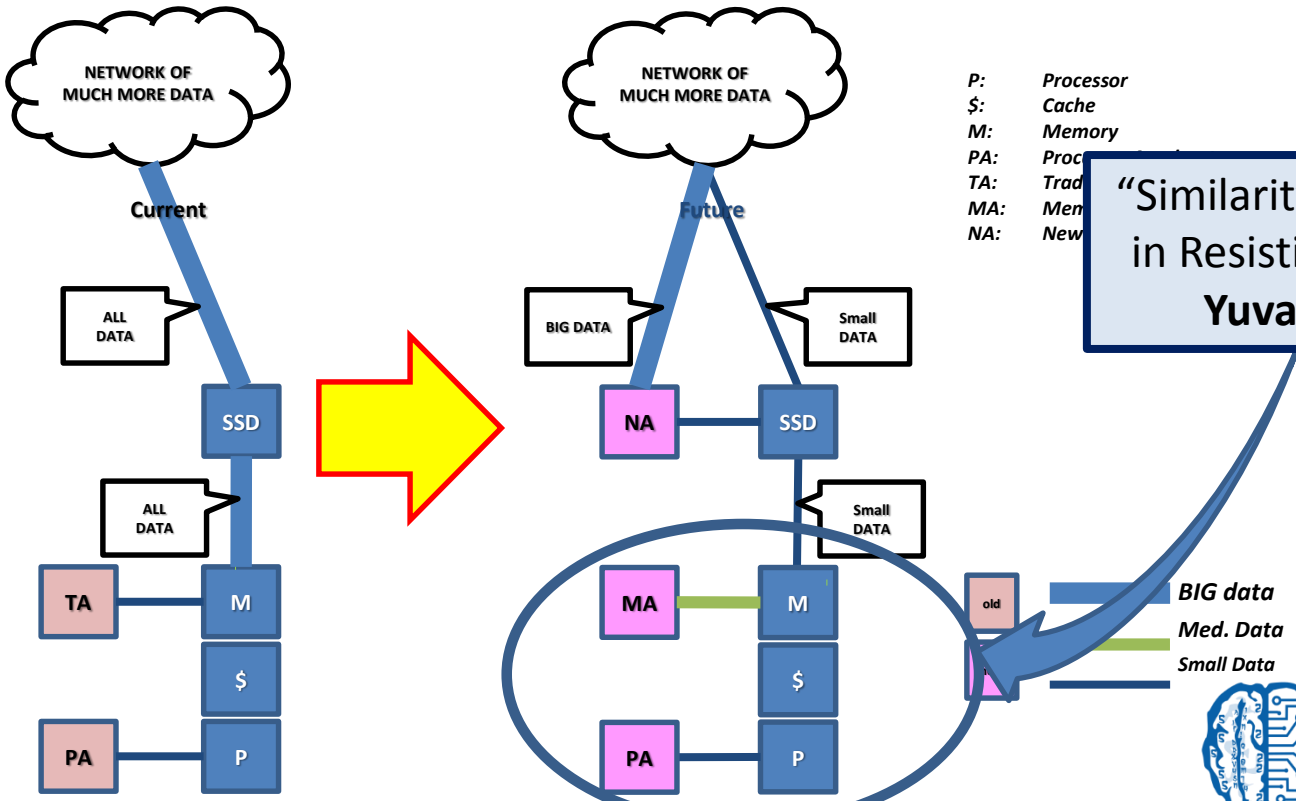
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence

Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. Data movement in/across nodes
2. Computation placed in the storage & network hierarchy
3. **New accelerators for big data**
4. **Applications and usage of new memory technologies (e.g. memristors)**
5. Leveraging ML algorithms for new microarchitectures



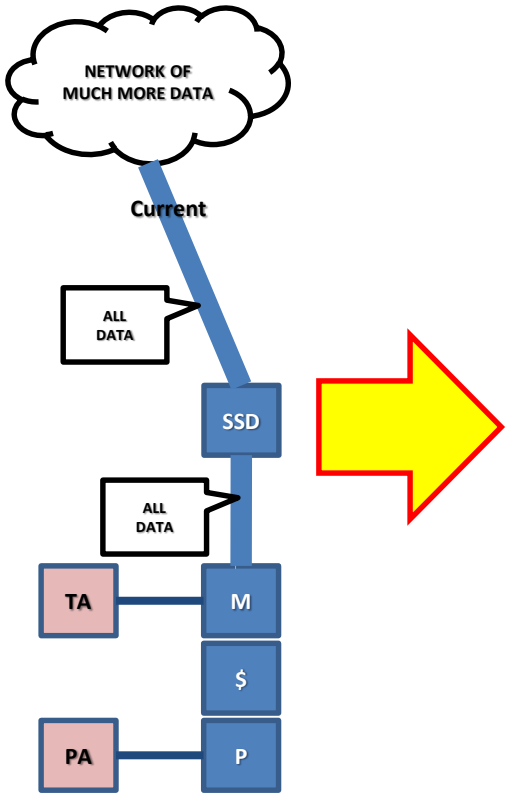
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

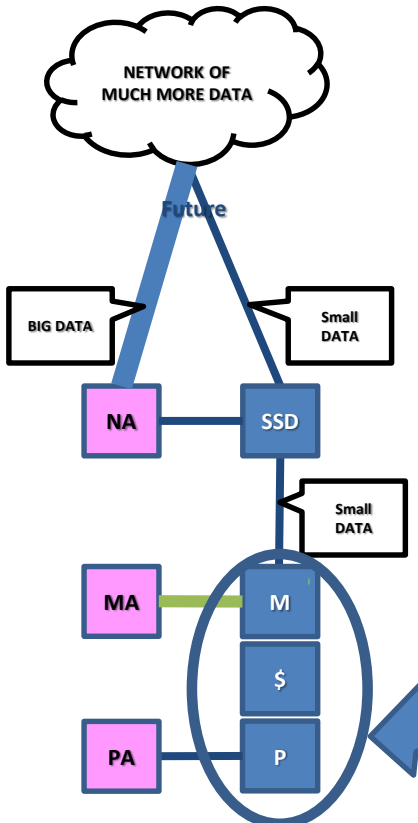
Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. Data movement in/across nodes
2. Computation placed in the storage & network hierarchy
3. New accelerators for big data
4. Applications and usage of new memory technologies (e.g. memristors)
5. **Leveraging ML algorithms for new microarchitectures**



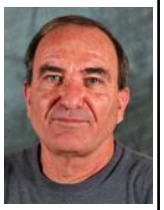


P: Processor
$: Cache
M: Memory
PA: Processor Accelerator
TA: Traditional Accelerator
MA: Memory Accelerator
NA: New Accelerator

"Using Neural Networks for Memory Prefetching…"
Yoav Etsion

ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

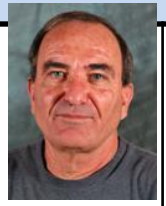
ICRI-CI Architecture Track – Year 4-5

Reduction of Memory traffic and solve Bandwidth System's bottleneck for Big Data		
<u>Funnel</u>:	*Identify System's Bandwidth issues in Big Data environment and suggest a remedy* 	**Prof. Uri Weiser**
<u>Accelerators</u> for Big Data & Machine Learning	*Novel Accelerators* 	**Prof. Ran Ginosar** Prof. Oded Schwartz
<u>Machine Learning for Architecture</u>	*Adaptive Systems* 	**Prof. Yoav Etsion** Prof. Uri Weiser Prof. Shie Mannor
<u>Memory Intensive Architecture:</u> **New memory based machine**	*New Technology based Architecture* 	Prof. Shahar Kvatinsky Prof. Avinoam Kolodny Prof. Eby Friedman (Technion/Rochester) **Prof Yuval Cassuto**

BACKUPS

ICRI-CI Architecture Track – Year 4-5

Update this foil

Reduction of Memory traffic and solve Bandwidth System's bottleneck for Big Data		
Funnel:	*Identify System's Bandwidth issues in Big Data environment and suggest a remedy* 	**Prof. Uri Weiser**
Accelerators for Big Data & Machine Learning	*Novel Accelerators* 	**Prof. Ran Ginosar** Prof. Oded Schwartz
Machine Learning for Architecture	*Adaptive Systems* 	**Prof. Yoav Etson** Prof. Uri Weiser Prof. Shie Mannor
		
Memory Intensive Architecture: **New memory based machine**	*New Technology based Architecture*	**Dr. Shahar Kvatinsky** Prof. Avinoam Kolodny Prof. Eby Friedman (Technion/Rochester) Prof Yuval Cassuto

Ongoing Interaction and Collaboration

Goal: Break-through performance and energy-efficien… …analytics platform

Time	Plan	Activit…
Q1'15	Education, background, select target & workloads.	Bi-wee… * Educa… * Lo…
Q2'15	Broad-stroke microarchitecture, performance, and workloads	* 5/… * 6/11/15 at H…
Q3'15	Next-level of detail for microarchitecture, performance, and workloads	Meetings: * DDIO system p…
Q4'15	High-level simulations and models of workloads on microarchitecture	Meetings: * 12/16/16Update … Weiser Funnel Research
Q1'16	Detailed simulations and models of workloads on microarchitecture	Meetings: * 2/17/16 Update on Ran Ginosaur Accelerators for Big Data Machine Learning * 3/9/16 Update on Shahar Kvatinsky Memory-Intensive Architecture
Q2'16	Detailed simulations and models of workloads on microarchitecture	Meetings: * 4/12/16 Update on Yoav Etsion Prefetcher using NN * 5/2/16 F2F at the Technion
Q3'16 -Q2'17	Parallel work on accelerators for target workloads and microarchitecture.	

Maybe not "detailed simulations and models" Use the words from the researchers' stated "major accomplishments"

ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel