Rapid-Rich Object Search Lab
博　云　搜　索　实　验　室

Architectures for Scalable Media Object Search

Dennis Sng
Deputy Director & Principal Scientist

NVIDIA GPU Technology Workshop
10 July 2014

NANYANG TECHNOLOGICAL UNIVERSITY

北京大學
PEKING UNIVERSITY

ROSE LAB OVERVIEW

Research Overview

Large Database of Media Objects

- Structured into multiple vertical application domains
- For machine learning & testing

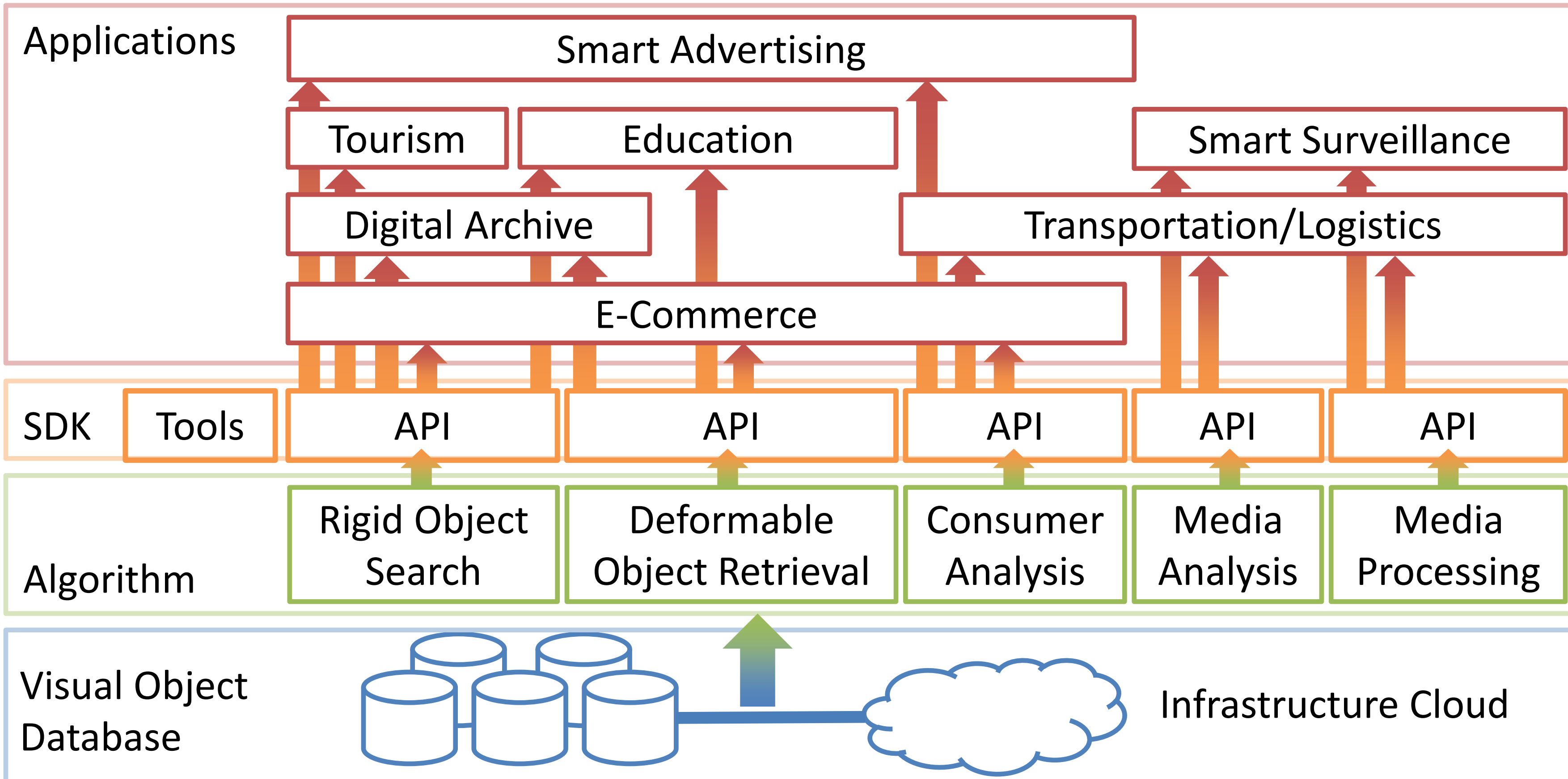
Next-Generation Object Search

- Fast & rich in content
- Real-time & contextual consumer behaviour analysis

Media Cloud Platform

- Scalable media search, processing, & delivery
- Testbed for experimentation

Solution Architecture



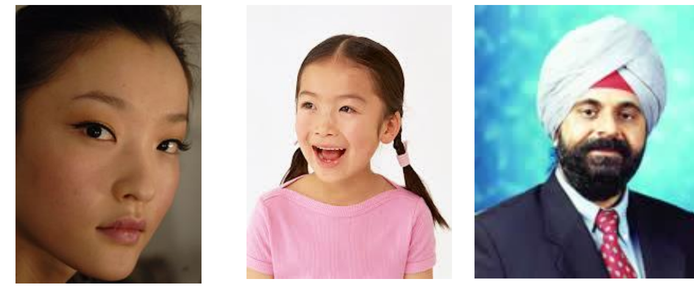
Object Categorisation

- **2D (Planar) objects:** Logos, book covers, CD covers, labels, coins

- **3D rigid objects:** Cars, hardware, product packages

- **Deformable objects:** Clothes, shoes, bags, toys

- **Faces:** Genders, age groups, profiles, ethnicity, sentiment



- **Landmark & Scenery**

ROSE Partner Ecosystem

Commercial Partners

Tencent 腾讯

inspur 浪潮

connex

Green Expo Technology

REACHFIELD IT SOLUTIONS
REACH IT WHERE IT MATTERS

Tigernix
We are users too

ADSC

Graymatics
Ultimate Multimedia Sensing

NTU/PKU Joint-Lab

Mobile Search with Contextual Mobility

Large-Scale Object Dataset & Analytics

Media Cloud Platform

Research Partners

北京大學 PEKING UNIVERSITY

Tencent 腾讯

Technology Partners

inspur 浪潮

Supported by

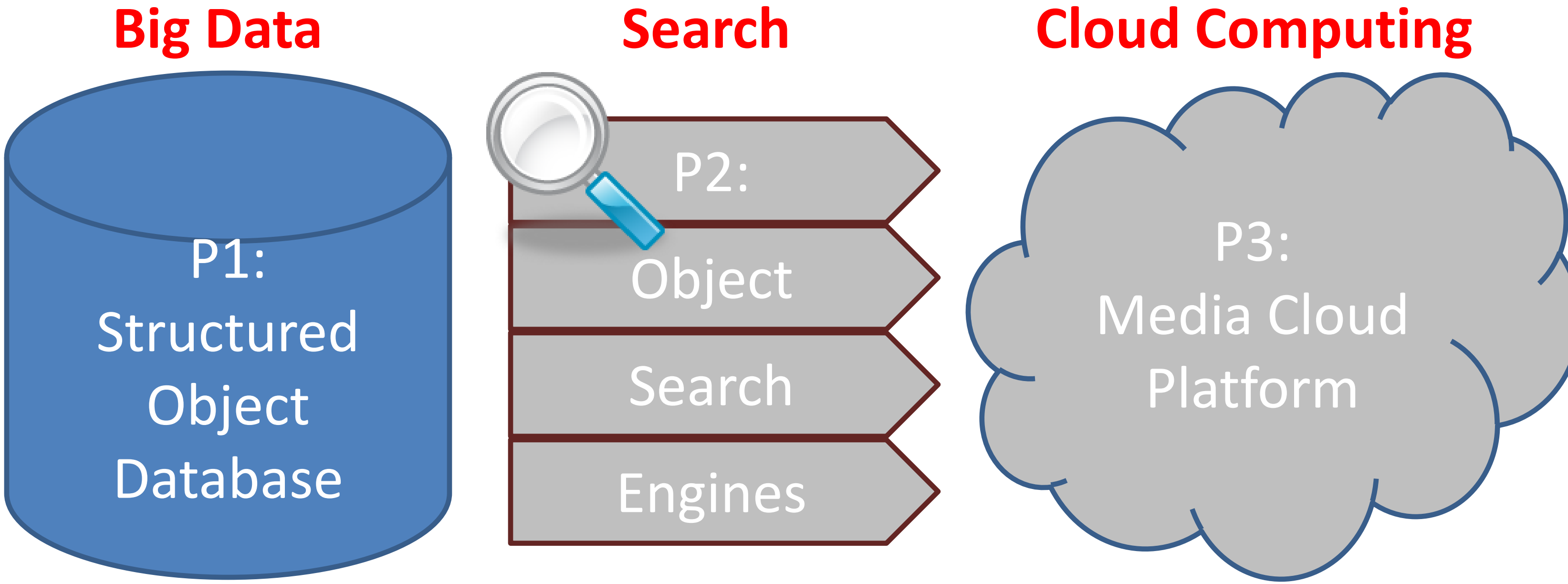
NATIONAL RESEARCH FOUNDATION
Prime Minister's Office, Republic of Singapore

mda Media Development Authority

Agency for Science, Technology and Research SINGAPORE

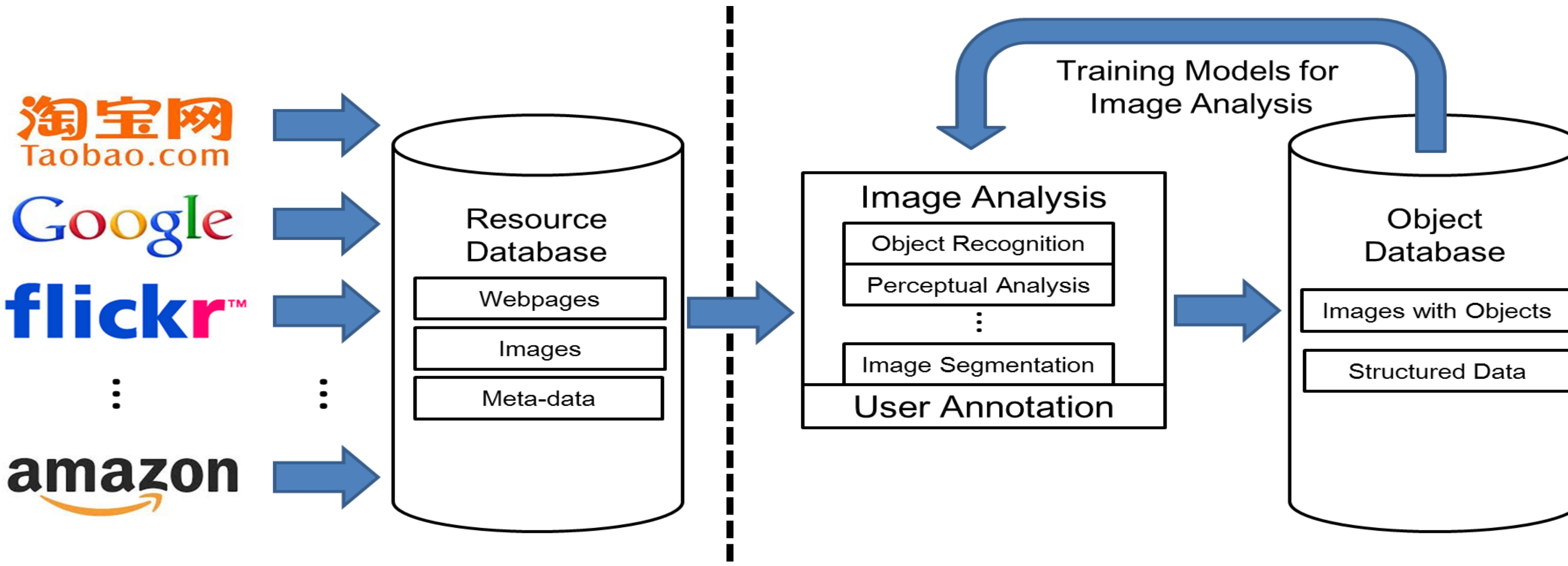
iDA SINGAPORE

Ministry of Education SINGAPORE



STRUCTURED OBJECT DATABASE

Framework: Object Database



Module 1 Crawler	Module 2 Resource Database	Module 3 Tools	Module 4 Object Database

Large Scale while High Quality

12/2013

- 220,000 raw images
- 170,000 non-identical images
- 45,000 clean images

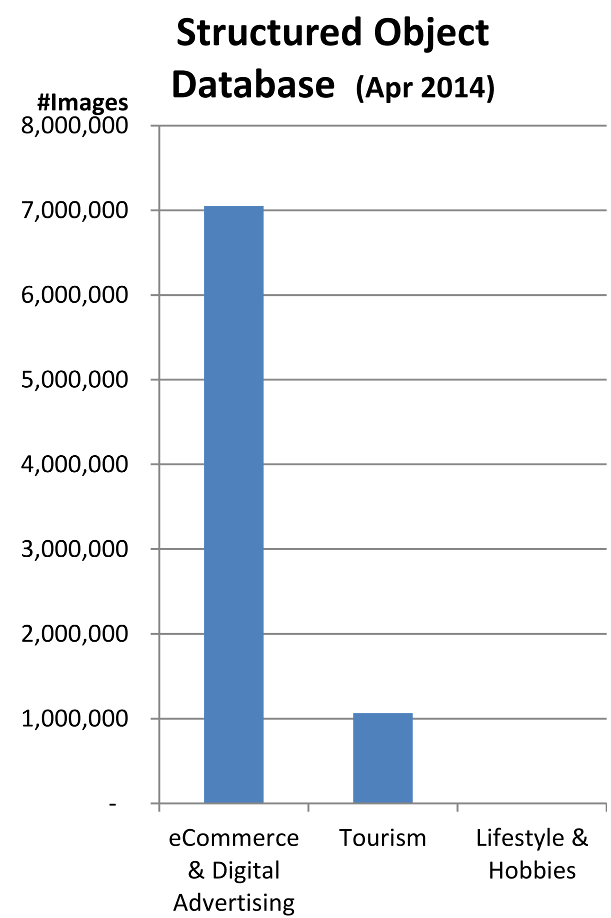
05/2014

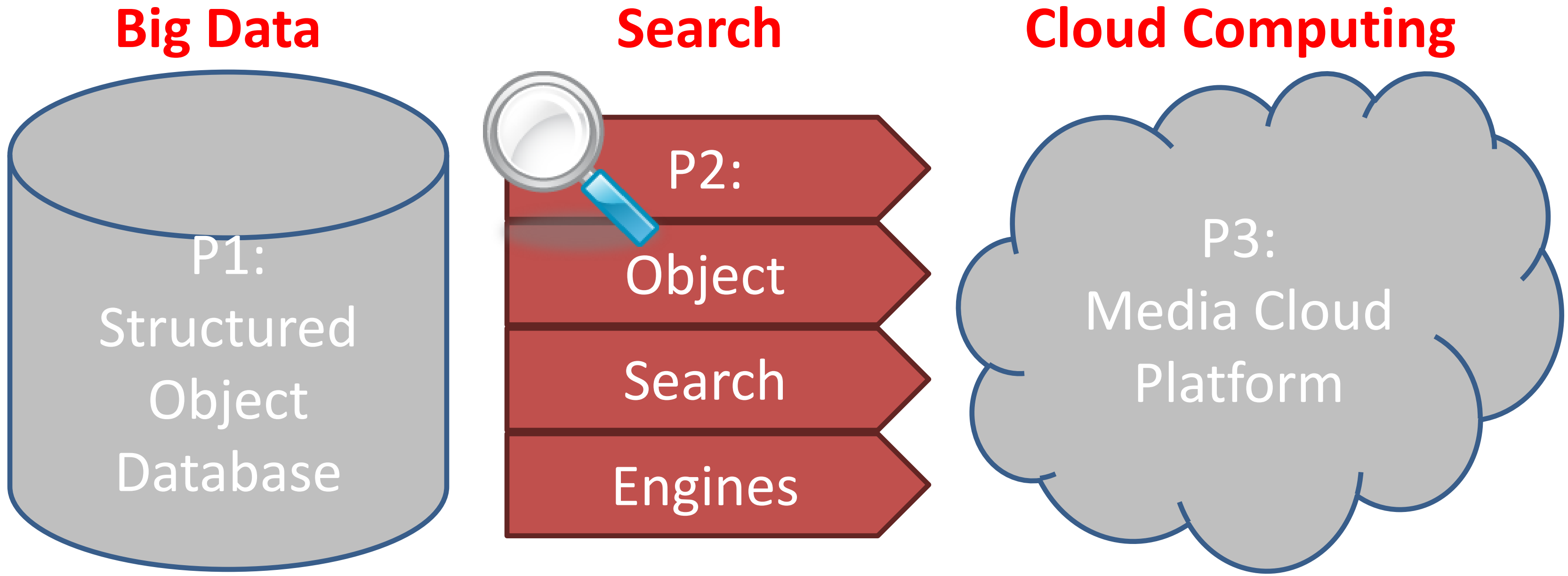
- 21 million raw images
- 17 million non-identical images
- 8 million clean images

Large-Scale Structured Object Database

50M structured objects (clean)

- **eCommerce & Digital Advertising**
  - Handbag
  - Shoes
  - Clothes
  - Book cover
  - Trademark & Logo
  - …
- **Tourism & Transport**
  - Landmark
  - Place of interest
  - Road signs
  - …
- **Lifestyle & Hobbies**
  - Fish
  - Fruit
  - Cat
  - Car
  - …





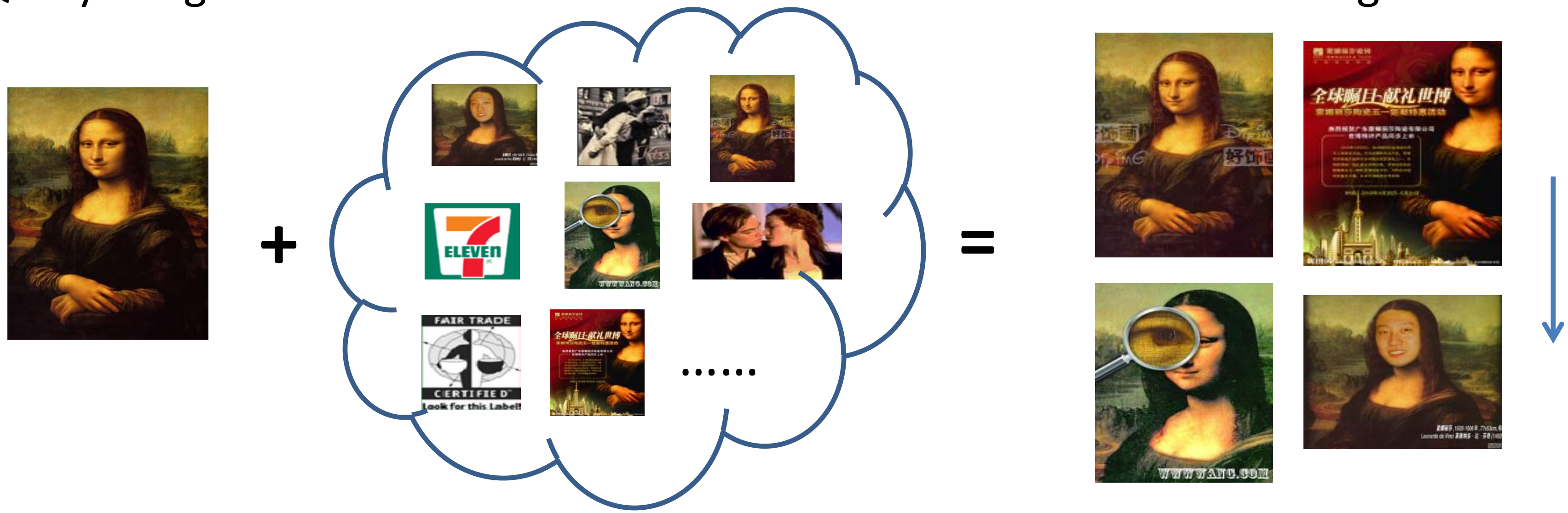
OBJECT SEARCH

Whole Image Retrieval

Query Image

Database

Ranked Images



Visual Object Search

Query Object

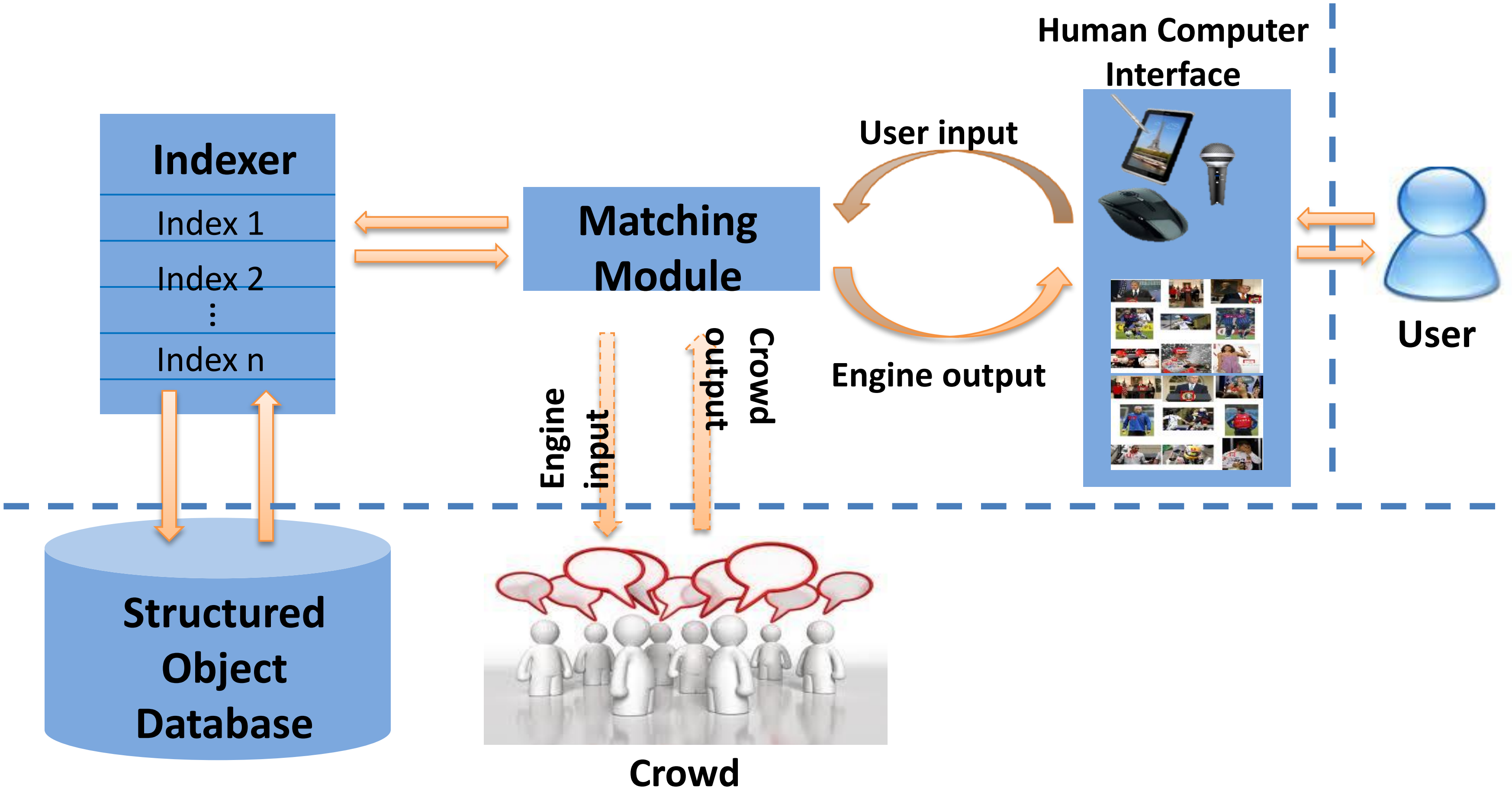
+

Database

=

Ranked Object Detections from Cluttered Images

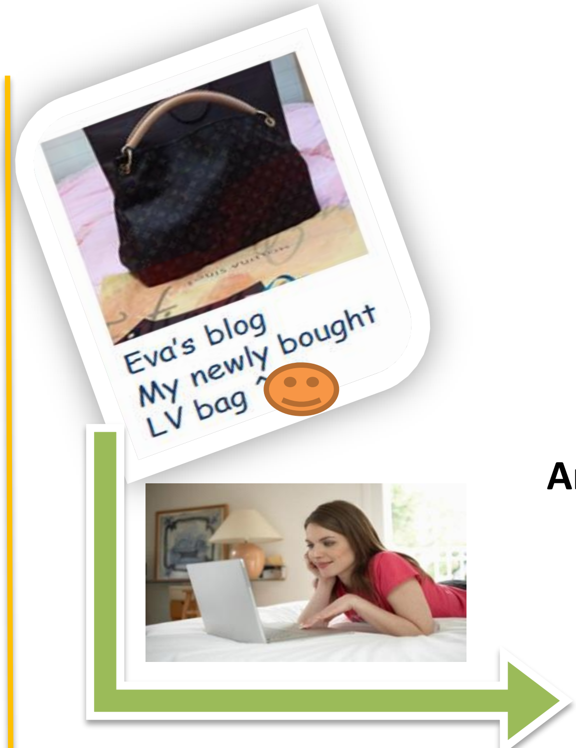
User Assisted Object Search Engine



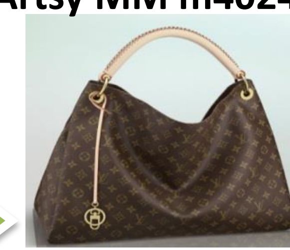
Branded Bag Recognition

- For people, use bag image to find bag

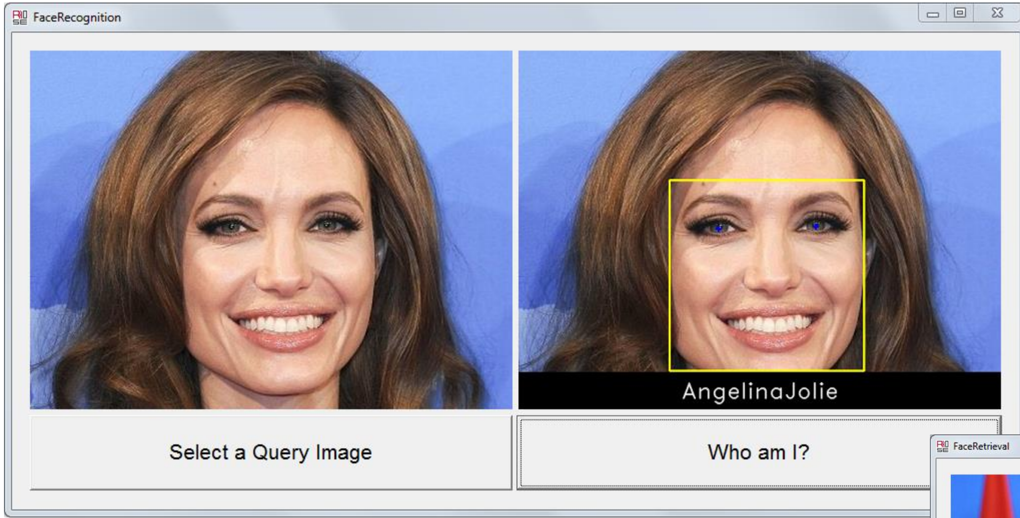
Street Scene



Artsy MM m40249

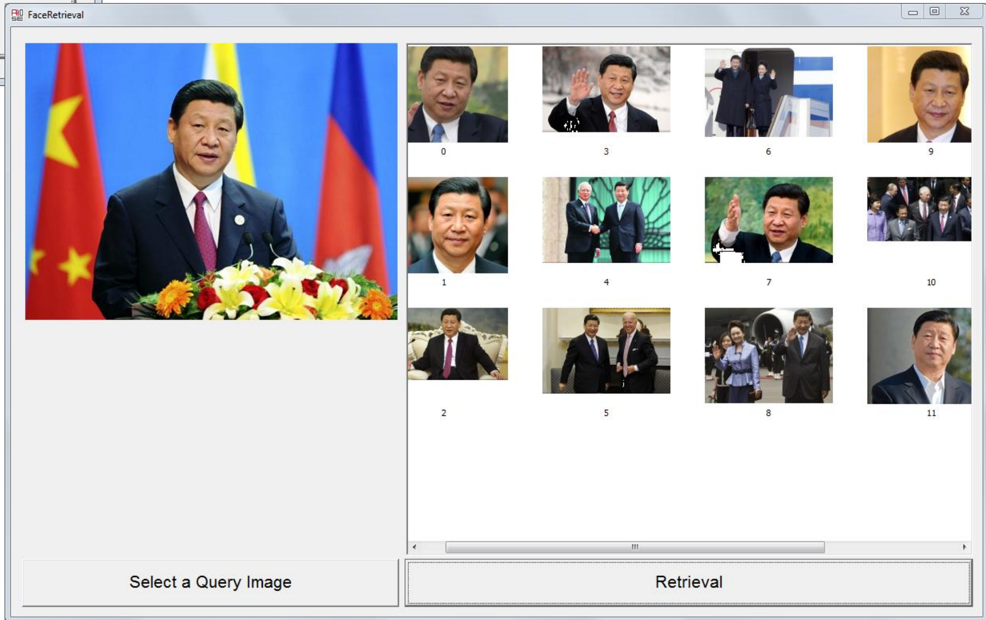


Face Recognition & Retrieval

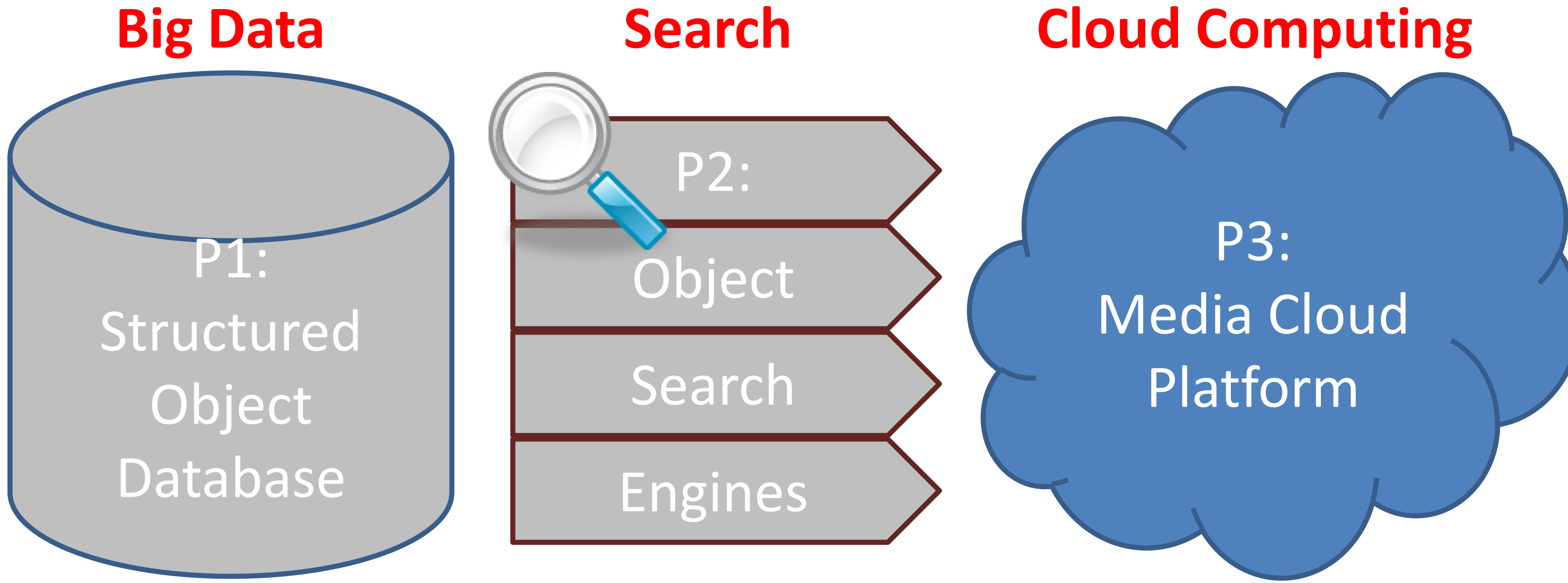


Identifying people

Retrieving images of a person

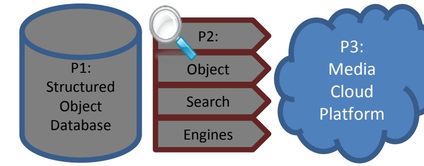


RIOSE Rapid-Rich Object Search Lab
博 云 搜 索 实 验 室



MEDIA CLOUD PLATFORM

P3: Media Cloud Platform



- Testbed
  - Design an innovative multimedia cloud platform as a test-bed for large-scale applications
- GPUs
  - For accelerating machine learning and object search
- Media Processing Technologies
  - Develop new media processing technologies in transcoding, visual analytics and quality assessment

ROSE Lab Physical Infrastructure

IT Cloud

- **P3 Cloud cluster**
  - 7x Dell R720 2U server without GPU
  - 84x Intel SNB Processor @2.3GHz
  - 434GB RAM @ 1600MHz
- **Network Infrastructure**
  - 1x CISCO Catalyst 3750-x Layer-3 Switch, 48 port (with 1 Gbps link to NTU Campus Network)
  - 5x CISCO Catalyst 3560-x Layer-2 IP-based Switch, 48 port

Experimental Zone

- **P2 Development cluster**
  - 1x HP Proliant 2U server
- **Experimental GPU Platform**
  - 3 x GPU Workstations: 2 x Titan Black/GTX770

HPC Cloud

- **GPU Cloud Cluster**
  - 1x Dell R720 2U server with 2x K20m GPU
  - 3x Dell R720 2U server with 1x K20m GPU
  - 1x Dell R720 2U server with 1x Intel Xeon Phil MIC

Big Data

- **P1 Database cluster**
  - 1x HP Proliant 2U server
  - 1x JBOD Storage Chasis
  - 1x D-Link NAS with 16TB storage
- **Storage Cluster**
  - 4x Novatte 12-Bay Storage Server
  - Up to 160TB Storage Capacity

ROSE Systems Infographic

Servers & Storage

- 23 Servers & Workstations
- 135TB Storage

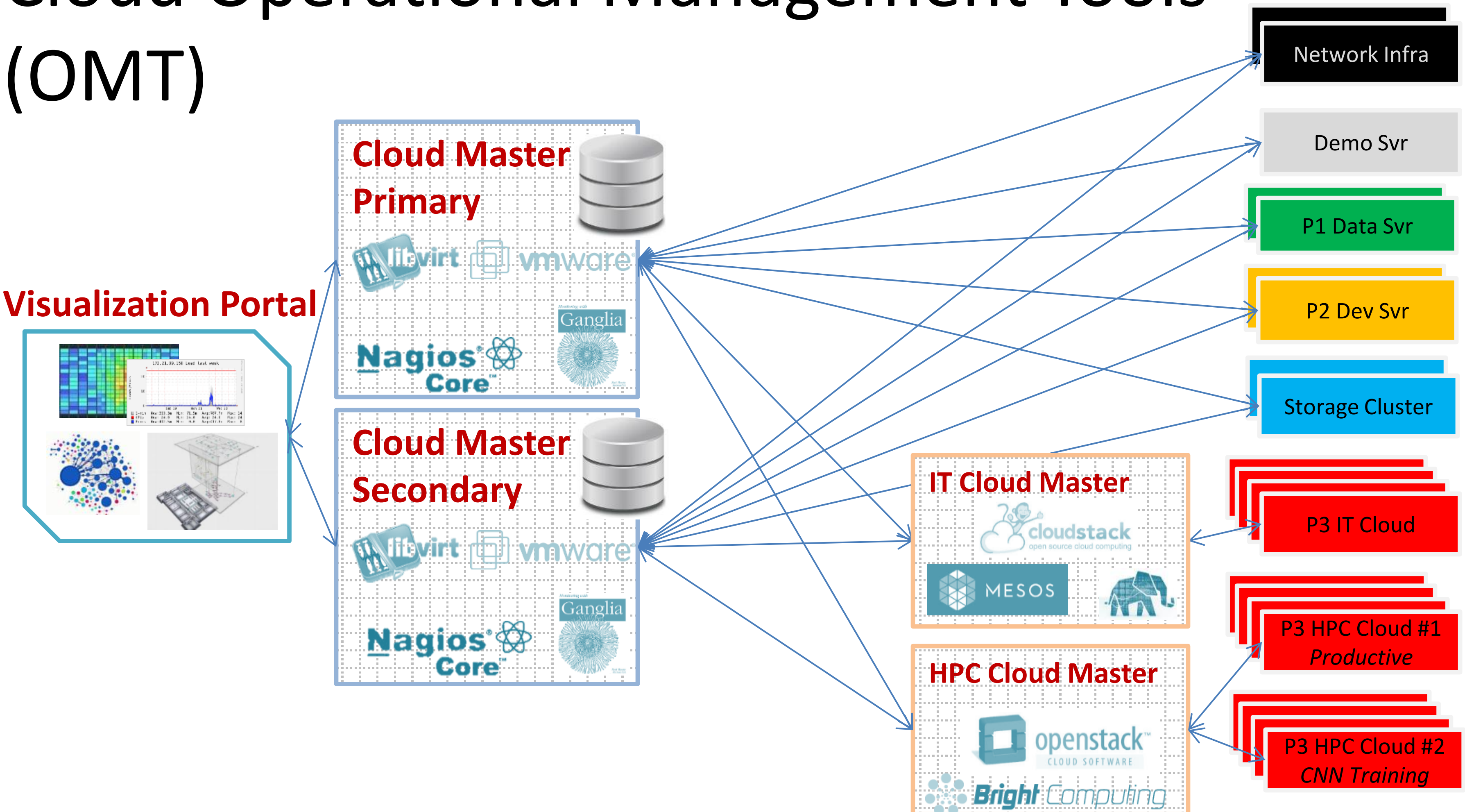
CPUs

- 41 CPUs
- 244 Physical cores
- 488 Hyper-threaded Logical cores

GPUs

- Includes Tesla K20 & K40, Titan Black, GTX770, GTX645,...
- 35 GPUs
- 36,288 CUDA cores

Cloud Operational Management Tools (OMT)



Deep Learning

ACCELERATING TRAINING TIME

Machine Learning

Google | Official Blog

Insights from Googlers into our products, technology, and the Google culture

Using large-scale brain simulations for machine learning and A.I.

Posted: Tuesday, June 26, 2012

g+1 1.7k Tweet 409 Like 568

Google Brain:
1,000 Servers
16,000 CPU-cores

Model: 1billion Connections
Dataset: 10million Images
Learning Time: 3 days

achieve significantly better accuracy. So we developed a distributed computing infrastructure for training large-scale neural networks. Then, we took an artificial neural network and spread the computation across 16,000 of our CPU cores (in our data centers), and trained models with more than 1 billion connections.

Nvidia GPUs Can Outperform Google Brain

By Kevin Parrish MARCH 26, 2014 2:02 PM - Source: Tom's Hardware US | 22 COMMENTS

TAGS : GTC 2014 GPUs Nvidia

Nvidia talks machine learning and the popular faces of humans and cats.

Nvidia's Jen-Hsun Huang talked a great deal about Machine Learning during his GTC 2014 keynote

NVIDIA GPUs:
3 Servers
12 GPUs
18,432 CUDA-cores

Cost: 100x less

that emulate how the brain functions. Our brains have neurons that recognize edges; we have a neuron for every type of edge. These edges turn into features that, when combined with other features, become a face. Computer scientists call this object recognition.

Deep Learning Model (1x K20)

- Date Set:

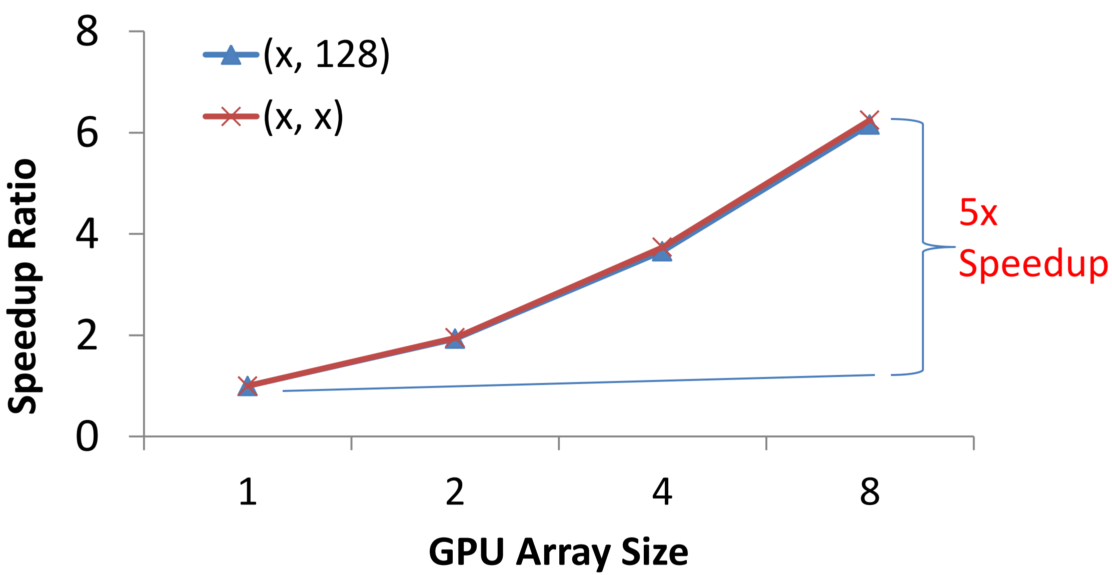
Dataset Name	#Images	#Category	Input Resolution
ILSVRC-2012	1.2 Million	1,000	224*224

- Model Scale:
  - No. of Fully Connected Layers: 3
  - No. of Convolutional Layers: 5
  - No. of Connections: 60 million
  - Size: 800+ MB

- Recognition Accuracy
  - 15.7% top-5 error rate

* ILSVRC = ImageNet Large Scale Visual Recognition Challenge

Speedup Result of Training same Model

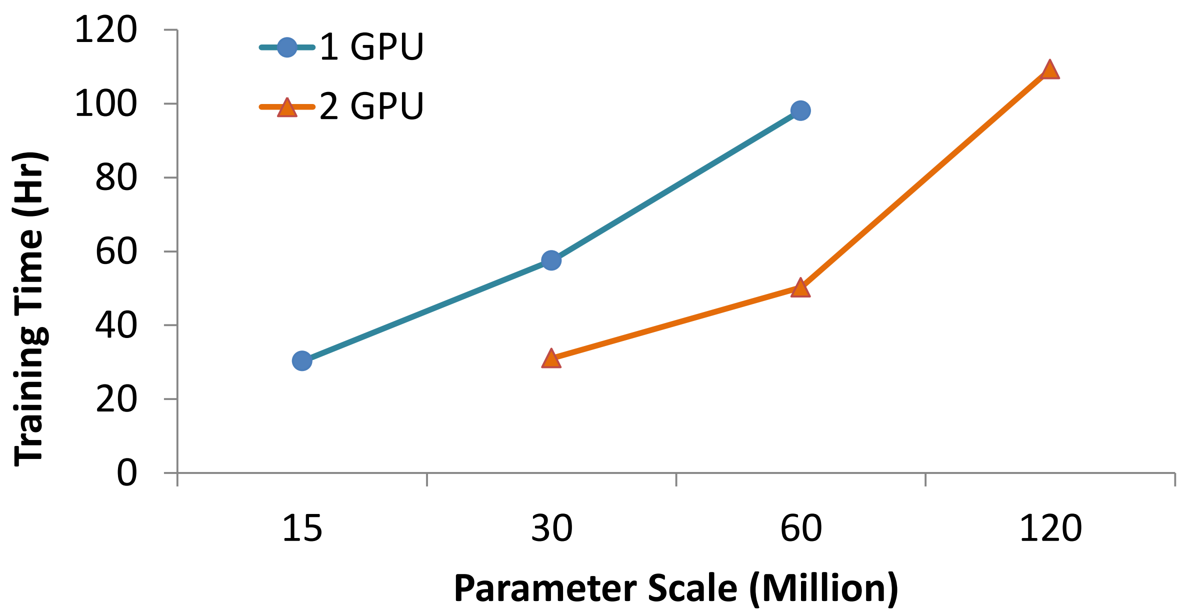


GPUs	Batch Size	Top-1 Error
1	(128, 128)	42.23%
2	(256 ,256)	42.63%
2	(256 ,128)	42.27%
4	(512 ,512)	43.58%
4	(512 ,128)	44.4%
8	(1024 ,1024)	43.28%
8	(1024 ,128)	42.86%

Top-1 Recognition Accuracy
(on par with competition winner)

Date Set: ILSVRC-2012

Training Time vs Scale of Model

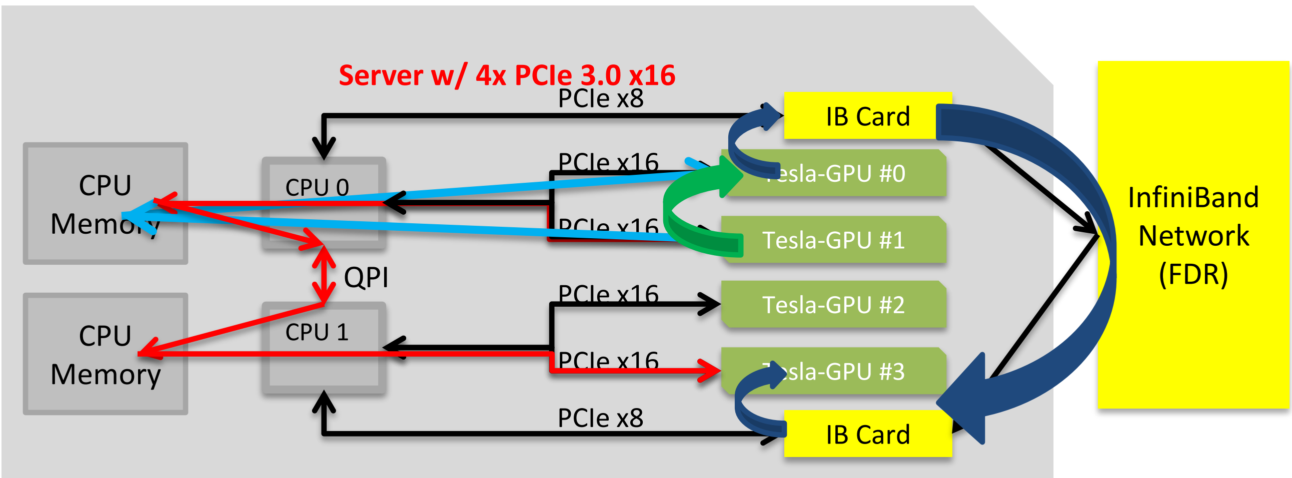


Date Set: ILSVRC-2012

Deep Learning

TRAINING PLATFORM REFERENCE ARCHITECTURES

GPU-GPU Communication Latency



(1) Without GDR P2P

- GPU to GPU DMA latency:
  - 2574.33 us (2MB DMA size)

(2) With GDR P2P

- GPU to GPU DMA latency
  - 524.28 us (2MB DMA size)

(3) With QPI

- GPU to GPU via QPI
  - > 2574 us (2MB DMA size)

(4) With GDR RDMA

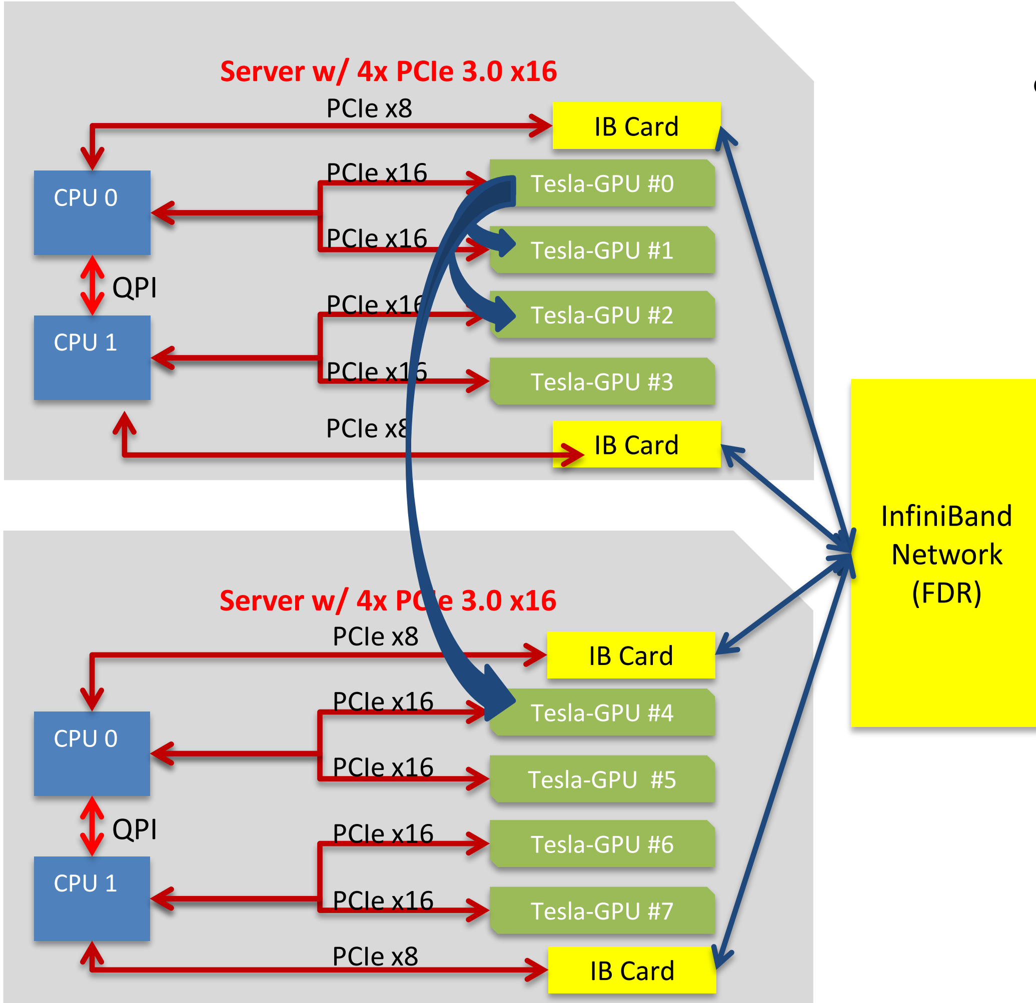
- GPU to GPU DMA latency
  - 600+ us (2MB DMA size)

Summary

- RDMA enables 4.9x Speed up!
- Cross-IOH DMA charges extra latency (60 – 70 ns)
- Cross-IOH DMA is not eligible to use GDR, latency > Without GDR P2P

* Based on the measurement released by GE

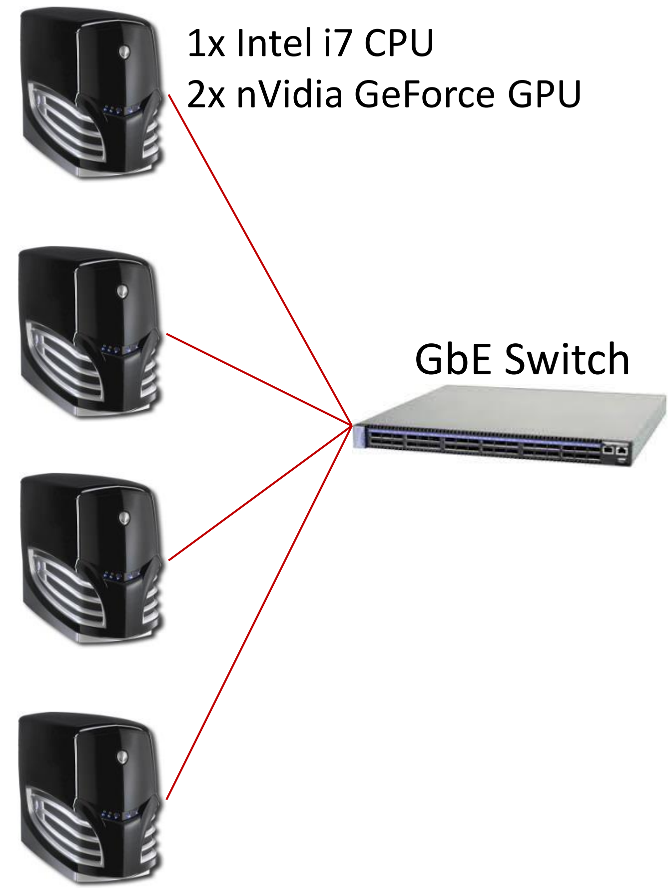
PCIe Layout & GPU-GPU RDMA



- Elementary Communication Models
  - Same Root Complex (eg. GPU-0 to GPU-1)
  - Same Server, Different Root Complex (eg. GPU-0 to GPU-2)
  - Different Server (eg. GPU-0 to GPU-4)

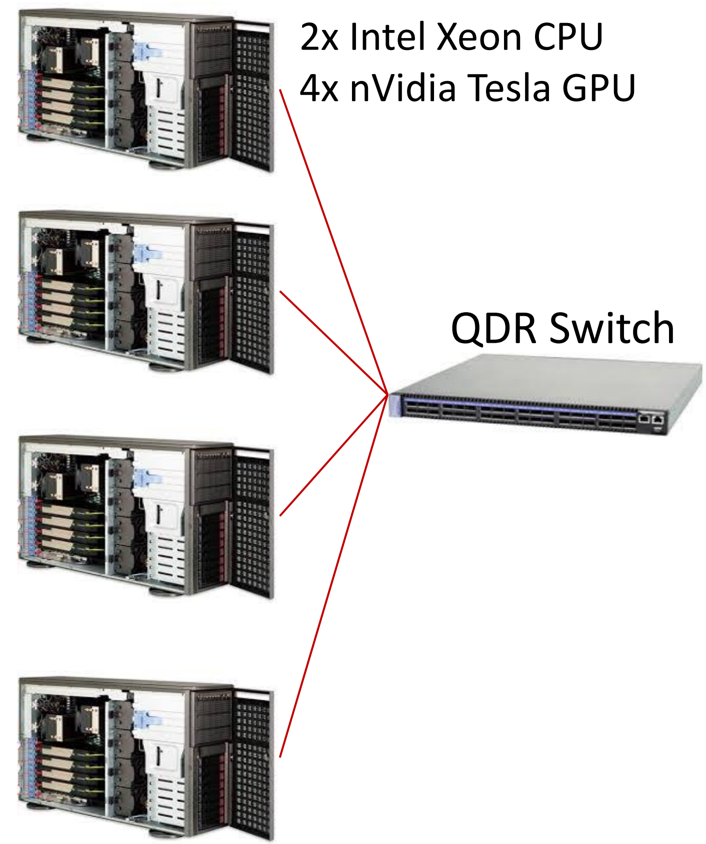
GPU Cluster with Commodity PC (for Development)

- Each node is High-end Commodity PC
- Nodes are interconnected via GbE network
- GPU communication using MPI



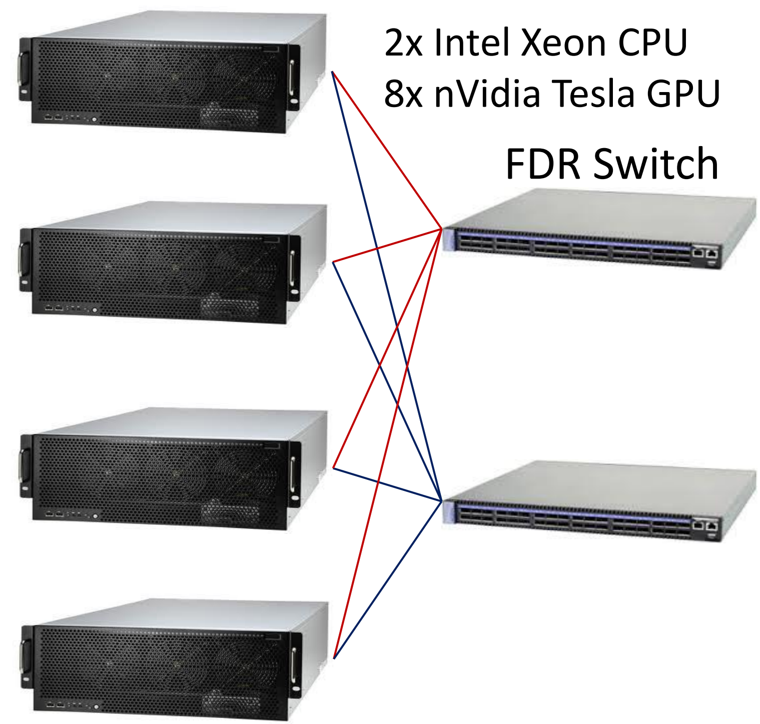
GPU Cluster with HPC Workstation

- Each node is High end Workstation
- Nodes are interconnected via IB network
- GPU communication using GPUDirect



GPU Cluster with HPC Server

- 4U Compute Node
- Nodes interconnected via multi-home IB network
- GPU communication using GPUDirect



Future Plans

- GPU Cluster as Deep Learning Training Platform
  - Various Inter-Connect Speed (eg. QDR vs FDR)
  - Various Inter-Connect Topology (eg. with & without redundancy)
  - Various GPU Processor (eg. K20 vs K40)
  - Various GPU Density (#GPUs per server)

- GPU-accelerated IaaS

- Deep Learning Training as a Service in GPU-aware Cloud

Industry Collaboration Models

- Research Programmes (Research Collaboration Agreements)
  - Covers ≥1 Joint Research Projects
  - Assignment of organisation's research staff to work with ROSE researchers
    - Can include Industrial Post-Graduate Programme (IPP) PhD students
- Technology Evaluation/Adoption Projects (Option Agreements)
  - Focus on evaluation of ROSE technologies, leading to licensing of the technology, OR
  - Focus on usage of Structured Object Database
- Affiliate Programme (Affiliate Agreements)
  - Newsletters, Briefings & Technology Demos for subscribers

ROSE Rapid-Rich Object Search Lab
博 云 搜 索 实 验 室

Thank You

rose.ntu.edu.sg/index.html

NANYANG TECHNOLOGICAL UNIVERSITY

北京大學 PEKING UNIVERSITY

Specs of Some Inter-connects

	Version	Frequency	Line Code	Single-Duplex per lane Bandwidth	Full-Duplex per lane Bandwidth	Single-Duplex Max lanes Bandwidth (GB/s)	Full-Duplex Max lanes Bandwidth (GB/s)	Original Transfer Rate	Small Message Minimum Interconnect Latency (< 64 Bytes)	Large Message Minimum Interconnect Latency @ 4194304 Bytes
QPI	NHL@2.4GHz	4.8GT/s				9.6	19.2		60 – 75 ns	
	???@2.93GHz	5.86GT/s				11.72	23.44		60 – 75 ns	
	SNB-E@3.2GHz	6.4GT/s				12.8	25.6		60 – 75 ns	
	IVB-E@3.6GHz	7.2GT/s				14.4	28.8		60 – 75 ns	
	???@4.0GHz	8GT/s				16	32		60 – 75 ns	
PCI-E	1.0	2.5GT/s	8b/10b	250MB/s	0.5GB/s	4	8	2.5GT/s		
	2.0	5GT/s	8b/10b	500MB/s	1GB/s	8	16	5.0GT/s	1.3 us	1251 us
	3.0	8GT/s	128b/130b	1GB/s	2GB/s	16	32	8.0GT/s	0.79 us	1072 us
	*4.0	16GT/s	128b/130b	2GB/s	4GB/s	32	64	16.0GT/s		
IB	QDR					5	10			
	FDR					7	14			