Clusters, Clouds, and Data for Scientific Computing 2014
La maison des contes, France

Future Dataflow-centric Computing for Streaming data from the Sky

Satoshi Sekiguchi

National Institute of Advanced Industrial Science and Technology
(AIST, Japan)

Collaborators:

Tomohiro Kudoh, Ryosuke Nakamura, Isao Kojima, and many others.

Abe Administration Statement （安倍内閣宣言）

『世界最先端ＩＴ国家創造』宣言 – June, 2013

- Statement on “Forging the World-Leading IT Nation”

Basic Philosophy （基本理念）

- Japan in Need of Revitalization （閉塞打破、再生日本）
- Toward the Achievement of the World-Class IT-Driven Society （世界最高水準IT利活用）

Abe Administration Statement （安倍内閣宣言）

Our Goals for the Future Society to create:

1. A society that enhances the creation of innovative technologies and new services and the growth of their providers
   - 増強革新的新産業・新服務的創作&全産業成長促進社会

2. The most natural disaster-proof society in the world, in order to ensure a safe and healthy lifestyle
   - 健康/安心/快適生活 & 社会抗災能力/安全的世界

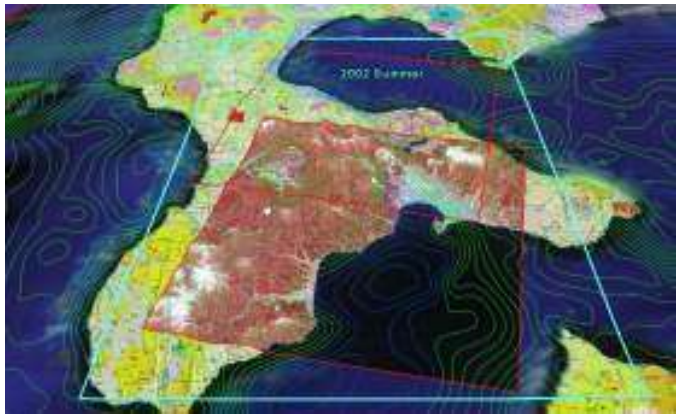
3. A society which guarantees that public services are available on a “one-stop” basis for anyone, anytime and any where

Efforts to Make to Fulfill the Goals

1. Creation of innovative technologies…
   a. Promotion to make the best use of both Open data and Big data
      - *Open data* for public, *Big data* for commercial use
2. The most natural disaster-proof society in the world to ensure a safe and healthy lifestyle
   a. Realization of a healthy and long-lived society through appropriate community-based health care, nursing-care support, and health promotion
      - Health monitoring (human and social infrastructure)
3. One-stop public services
   a. Innovation of the government information system across both local and national government
      - “Cloud First” Concept for the system cost reduction

Objectives of the GEO Grid (since 2005)

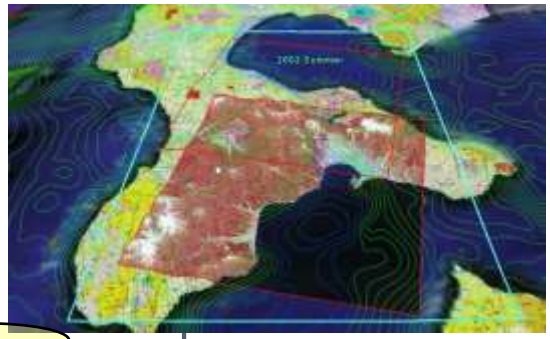
- Help Geo-* scientists to understand
  - Global warming, inventory of carbon dioxide
    - Kyoto protocol, environmental burden
  - Alternate energy
    - Biomass
    - Wind-power generator network
  - Harvest yield prediction/estimation
    - Weather, Soil, temperature, humidity, sunshine, etc.
- Help decision makers to plan
  - Hazard mitigation
    - Earthquake, Landslide, Flood, Volcano eruption, Tsunami
  - Exploration of natural resources
    - Oil, natural gas, mineral
- Unbeknown applications
  - Games, Amusements, Personal geo-record/history, etc.
  - Social science apps



A Workflow example

"Disaster prevention and mitigation (Landslide)"

Geology map (GSJ)



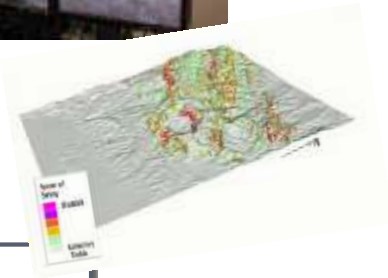
Rain fall
(AMEDAS/Fieldserver)

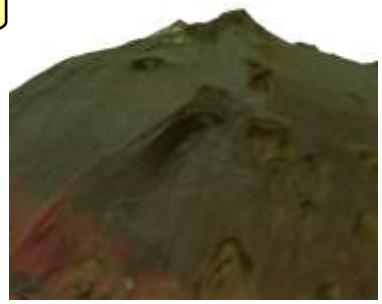
Early warning system
based on Susceptibility map

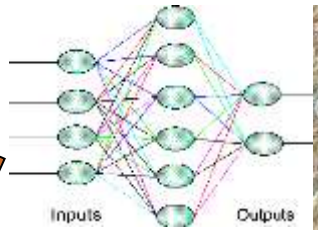
GEO

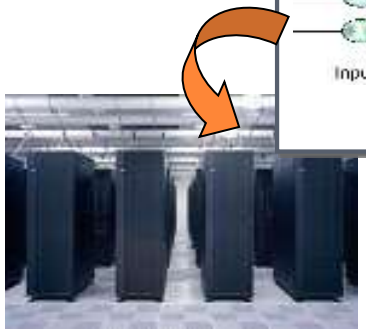
MET

ASTER







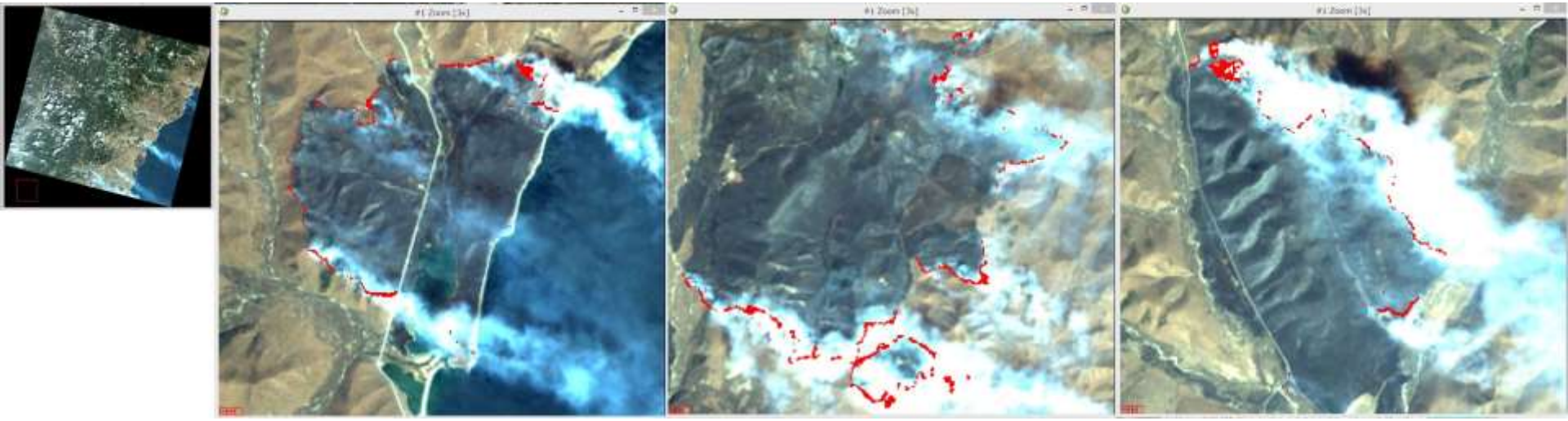


High resolution DEM
provided from ASTER

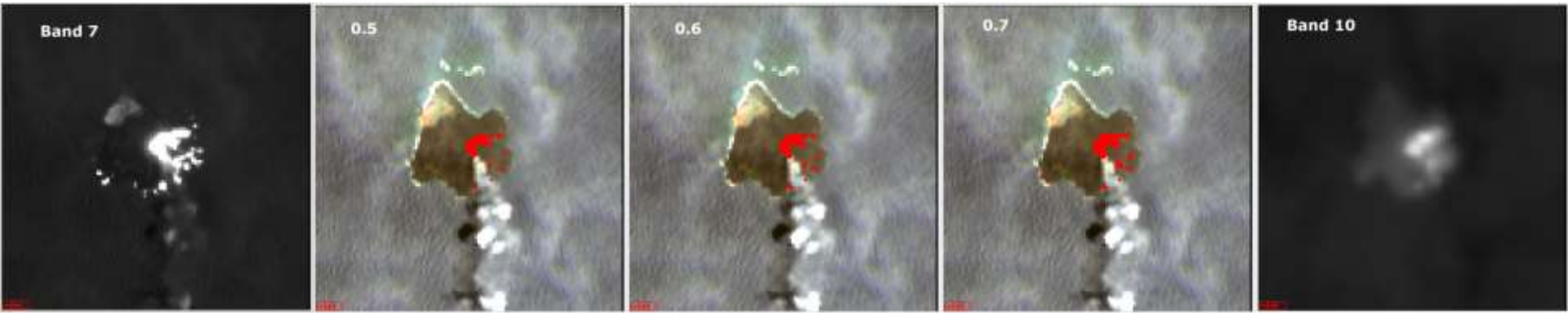
Large scale Computer simulation
using actual landslide DB

ASTER hot spot detection system

Wildfire in Russia

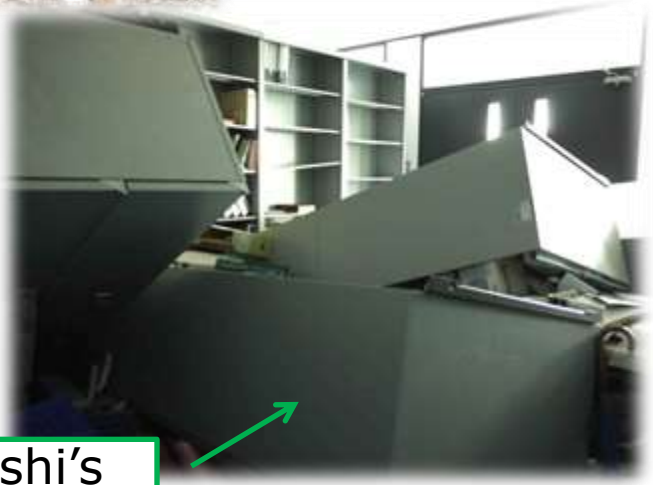


Volcano on Nishinoshima island



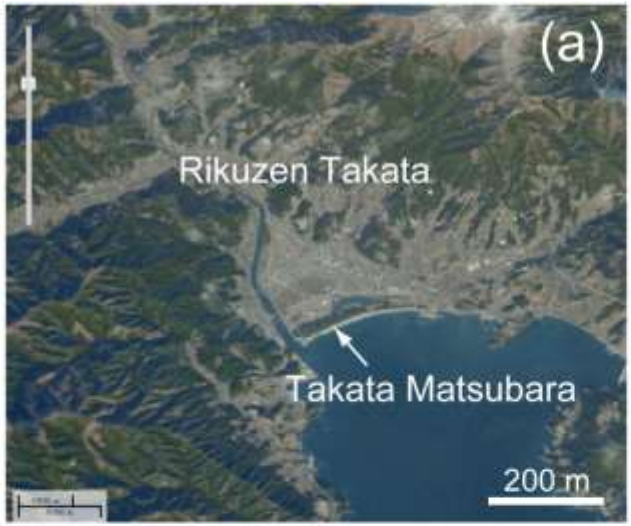
GEO Grid for Disaster Response

AIST's machine Rm



Satoshi's office

ASTER Comparison Before and After Tsunami



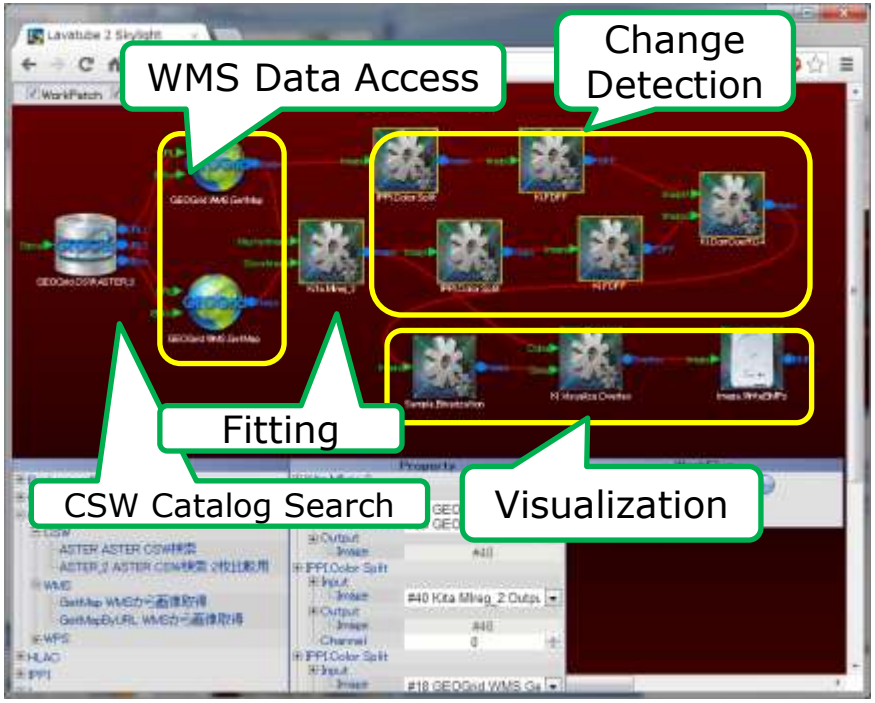
(a) 2008/12/25
(b) 2011/3/14
(c) 2011/3/19
(d) 2011/3/28

Rikuzen-takada

The tsunami reached a very deep part of the bay and changed the shape of the coastline.

Example: Change Detection Application of GEO Grid

1. Metadata Search(OGC CS-W) for a specific area
2. Data Acquisition(OGC WMS)　for 2 satellite images
3. Position/Color Fitting　of satellite images
4. Change Detection

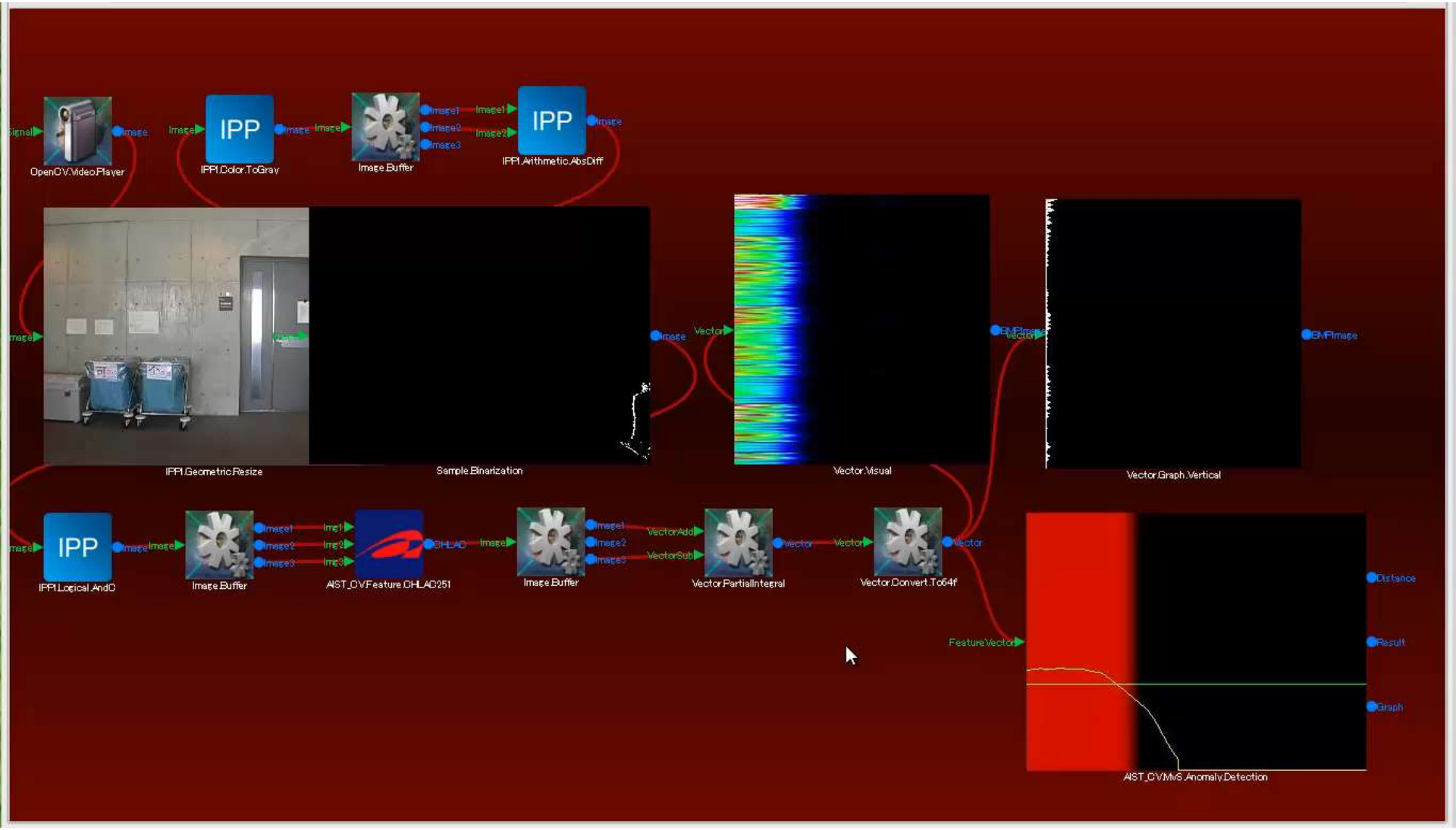


Workflow constructed with Lavatube

2011.06

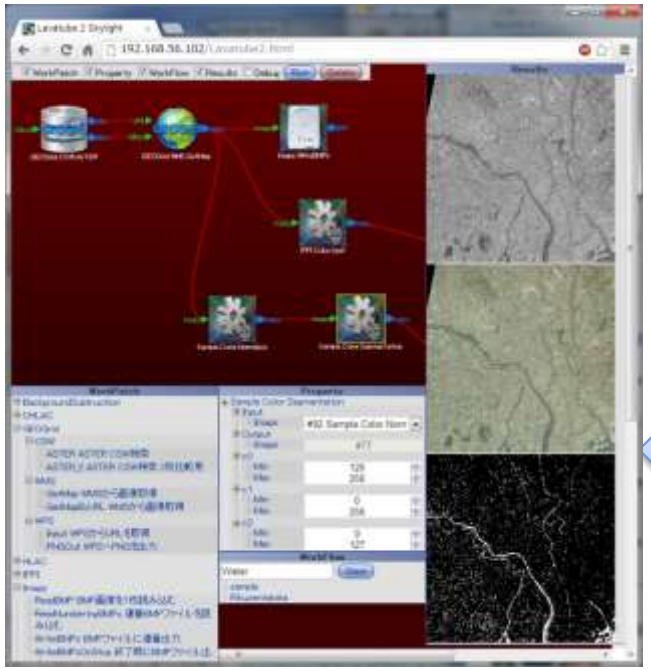
2012.06

Result

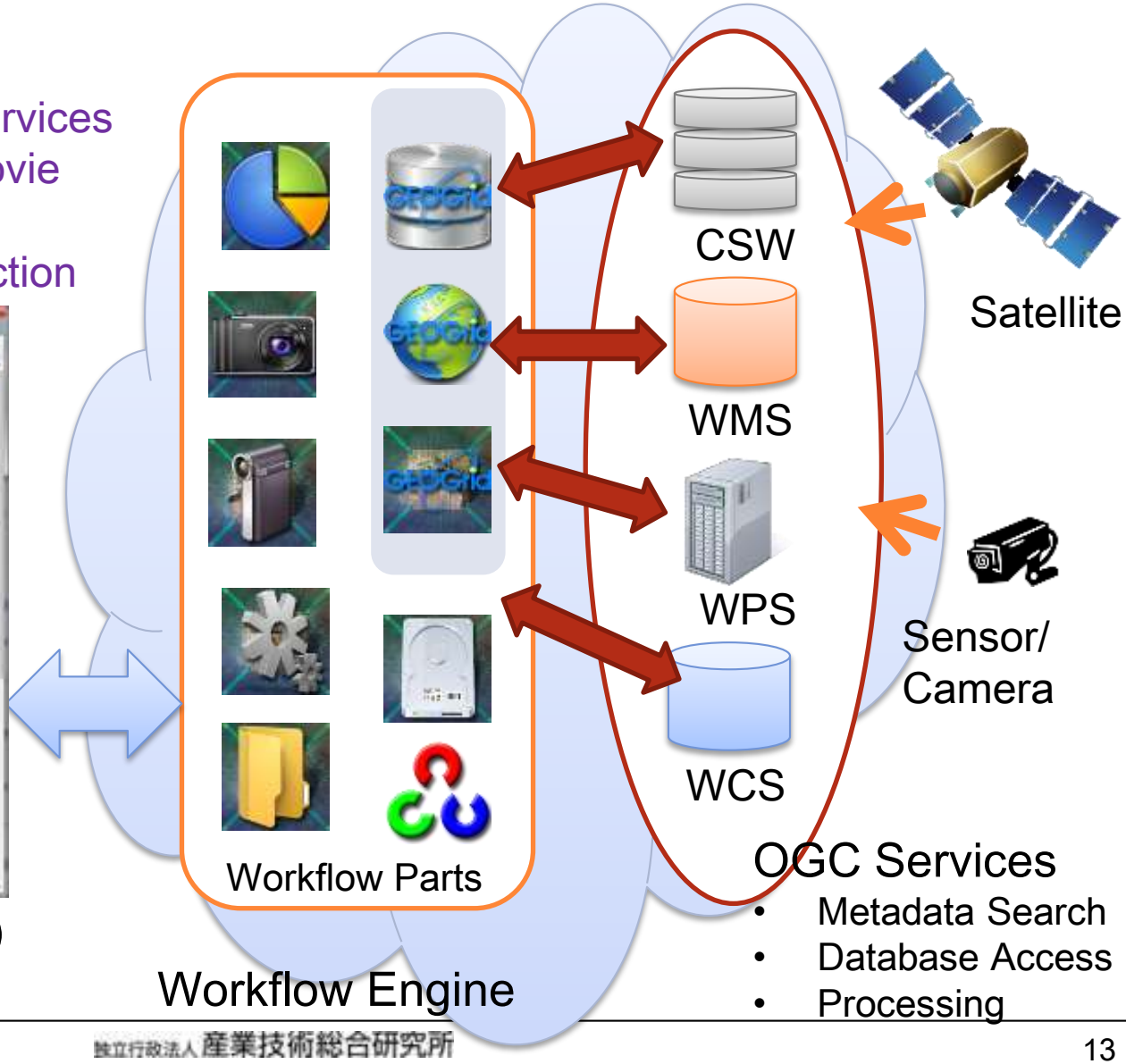

OpenCV.VideoPlayer
IPP
IPP.Color.ToGray
Image.Buffer
IPP
IPP.Arithmetic.AbsDiff
IPP.Geometric.Resize
Sample.Binarization
Vector.Visual
Vector.Graph.Vertical
IPP
IPP.Logical.AndC
Image.Buffer
AIST_CV.Feature.CHLAC251
Image.Buffer
Vector.PartialIntegral
Vector.Convert.To64f
FeatureVector
Distance
Result
Graph
AIST_CV.MvS.AnomalyDetection

Lavatube for spacio-temporal data and image/movie processing

1. Support rest-based OGC (OpenGeospatial Consortium) services
2. Support various image/movie processing modules
3. Provide High-Level interaction

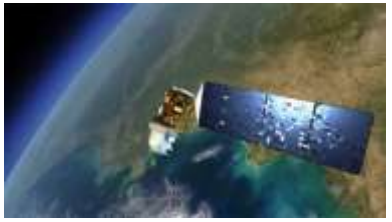


Browser Interface(HTML5)
or Windows engine

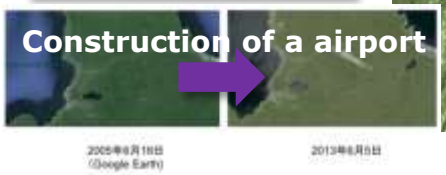


New Data Sources!

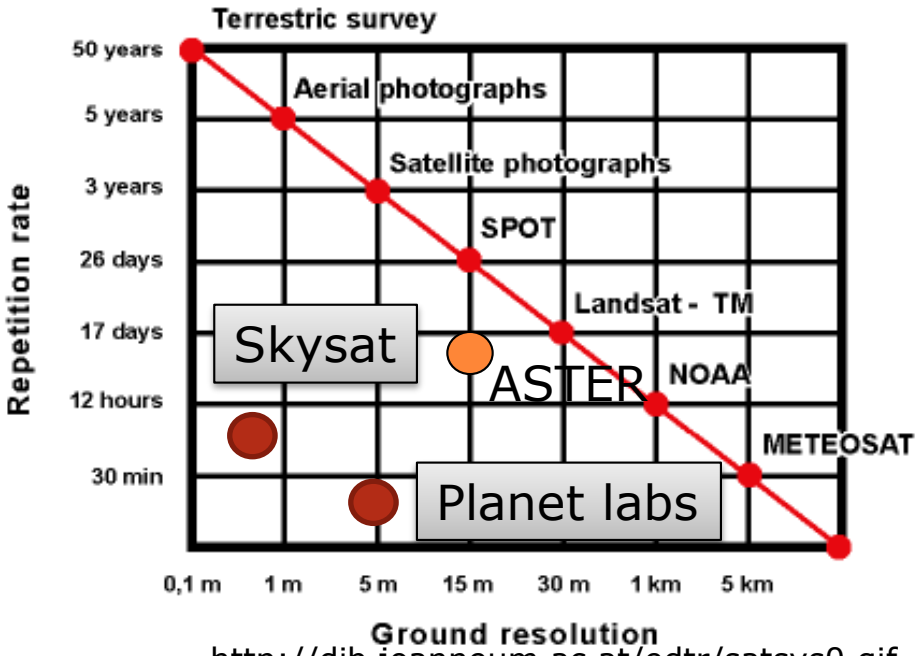
- Landsat-8 (by USGS)
  - Latest earth observation satellite launched 2013.
    - 15m/Pan 30m/Color
    - 16days cycle for the same area
  - Free and Open!



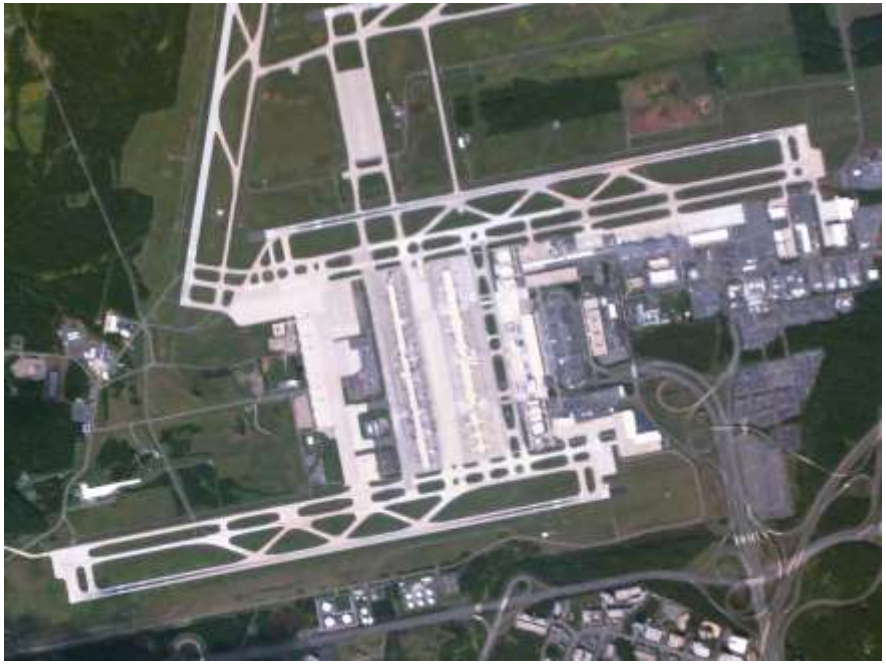
- AIST set up the ground station for Landsat-8(with Tokai-U)
  - Daily data directly from the satellite
  - Upload the images in the Internet < 2hr
    - Thanks to HPC (c.f.1 day in USGS)



Game Changers: simple sat – low orbit – high resolution

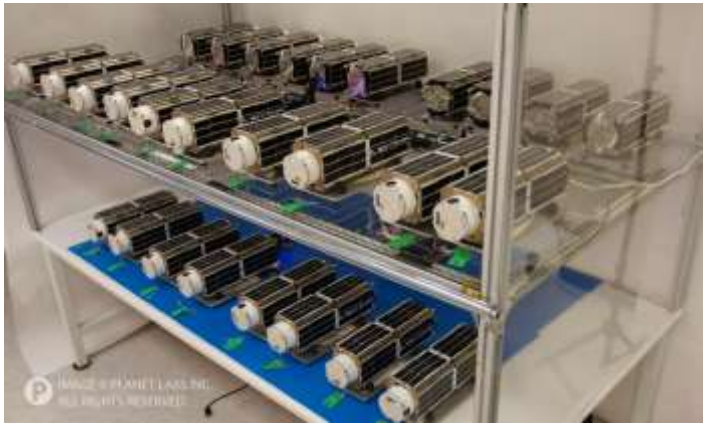


http://dib.joanneum.ac.at/edtr/satsys0.gif



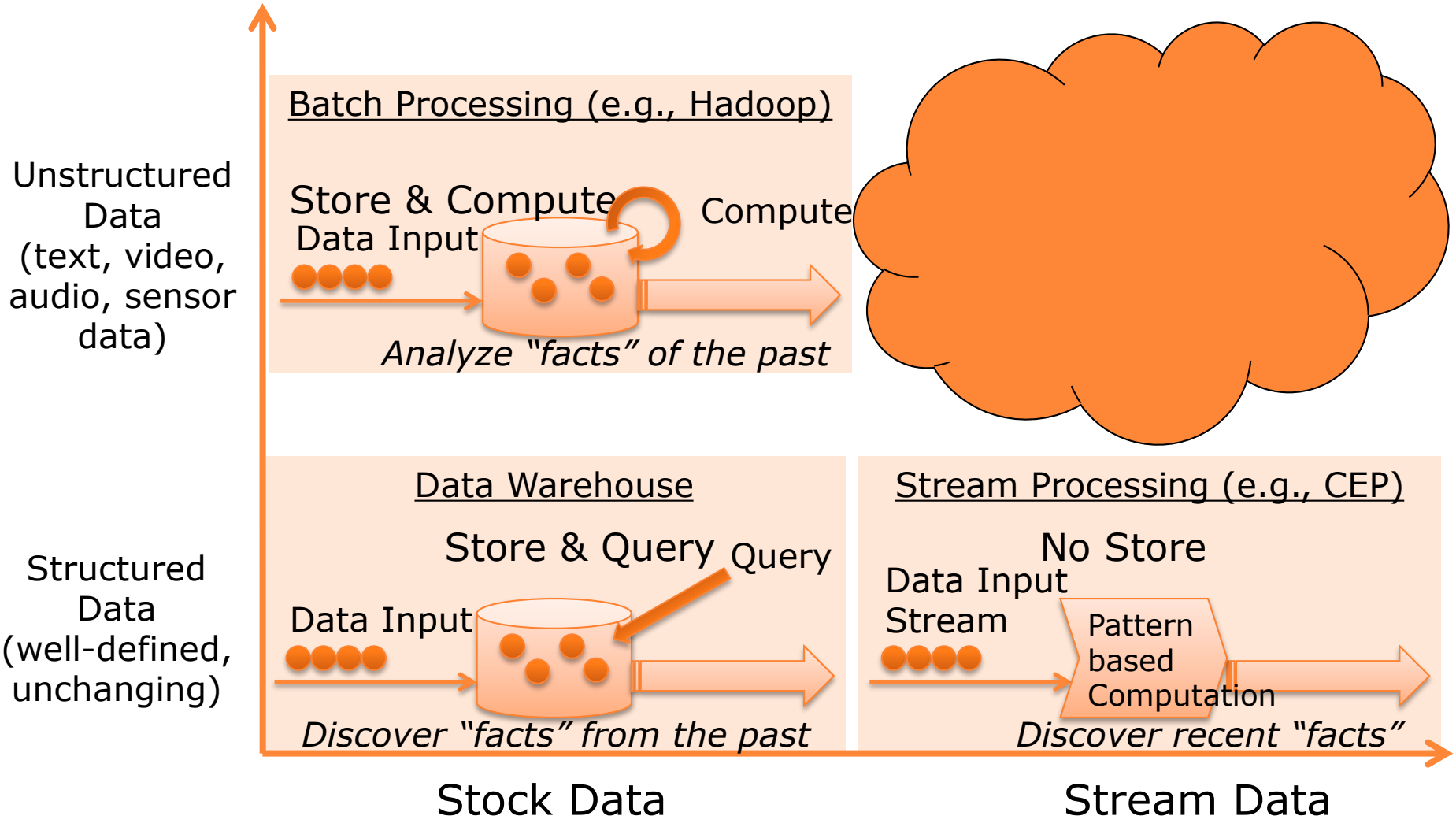
https://www.planet.com/gallery/images/dulles-international-airport.jpg

http://www.firstimagery.skybox.com/2014/7/10/port-au-prince-haiti



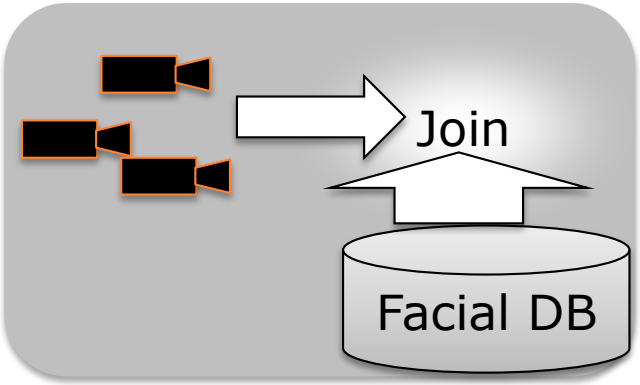
https://www.planet.com/flock1/

Big Data Challenges

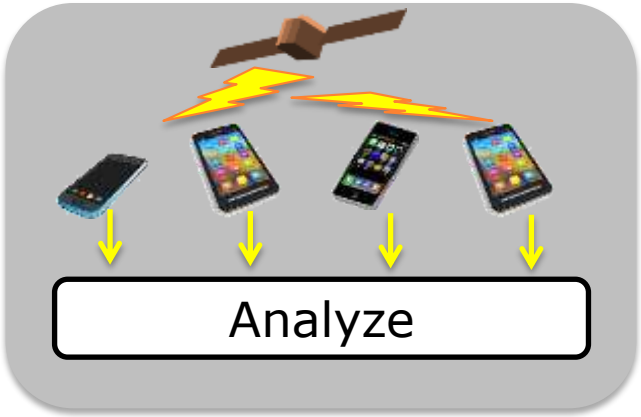


Real-time Big Data Processing in 2020

- More data will be processed to get better results
  - Video Monitoring
    - # of cameras = 0. 1 M
    - Joining with Facial feature vector
      1 KB x 1Billion (1 TB)



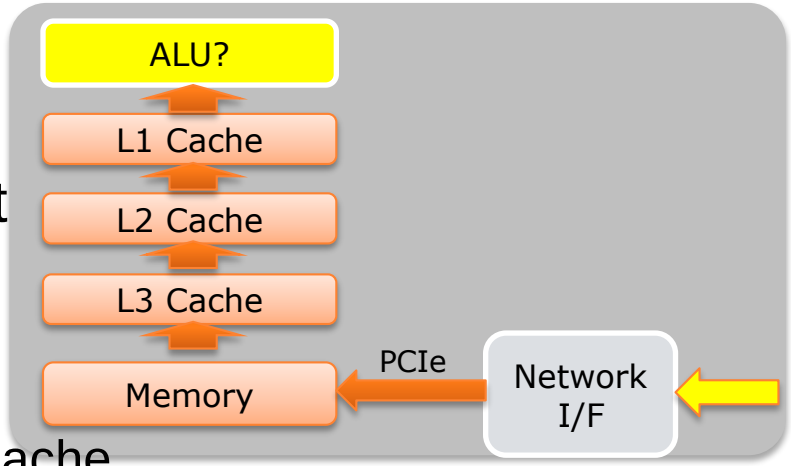
  - Smartphone GPS Data Analysis
    - # of smartphones = 100M
    - Car / Train / Pedestrian Traffic analysis, Traffic jam prediction



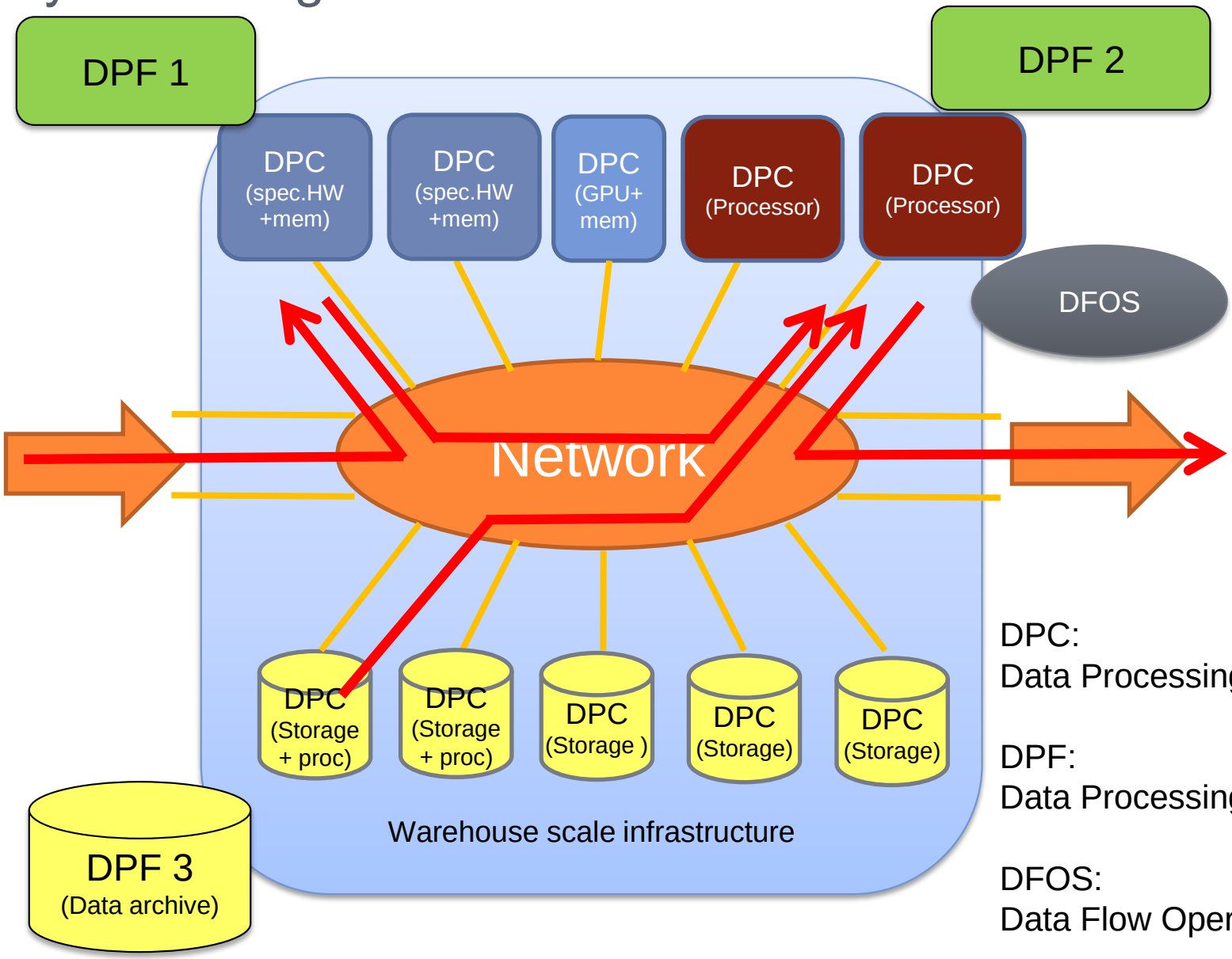
- Multi-modal Analysis
  - Analyze Tweet stream with video stream
  - Analyze Tweet stream with smartphone stream

Issues in Current Systems

- Required performance for future real-time big-data applications cannot be achieved
  - Usually, independent OS runs at each server, and servers are coordinated by an upper layer middleware
    - Large overhead caused by unnecessary processing in OS
    - “OS jitter” degrades the system performance
  - Data transfer and processing are not tightly coupled
    - Poor performance for communication intensive applications
    - Processing performance depends on cache memory. Data transfer to/from cache is a bottleneck.



System Image



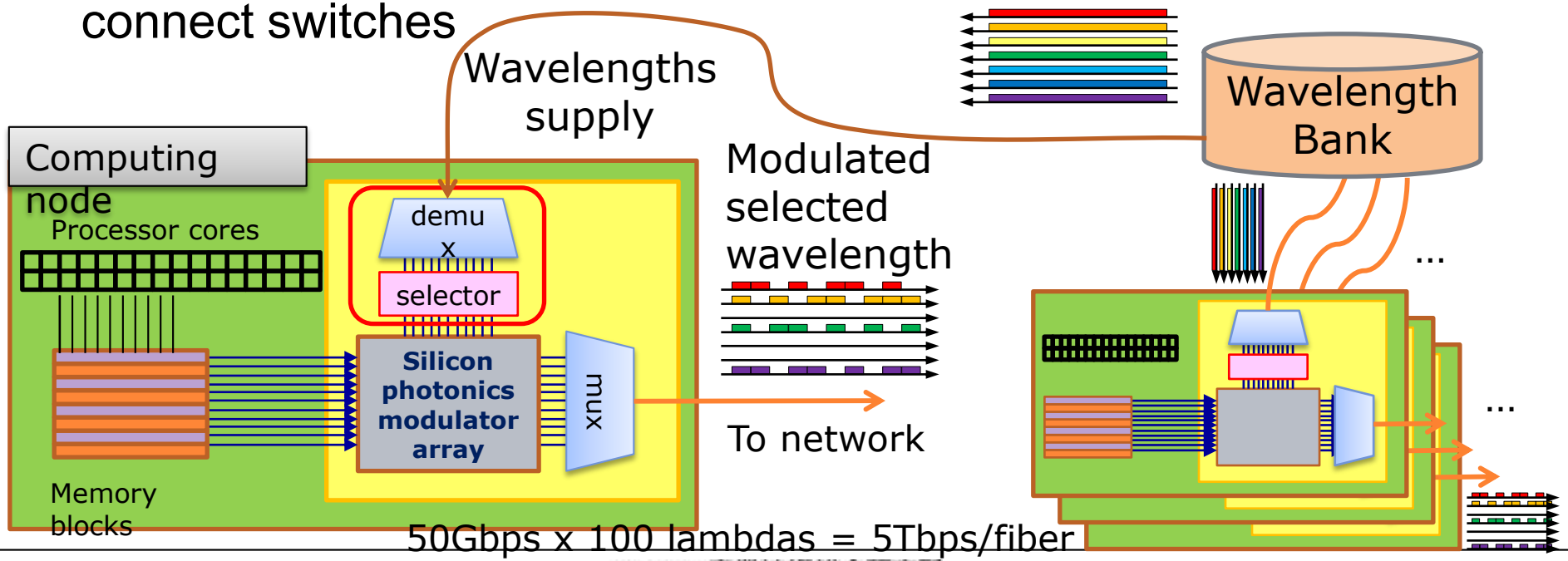
DPC:
Data Processing Component

DPF:
Data Processing Function

DFOS:
Data Flow Operating System

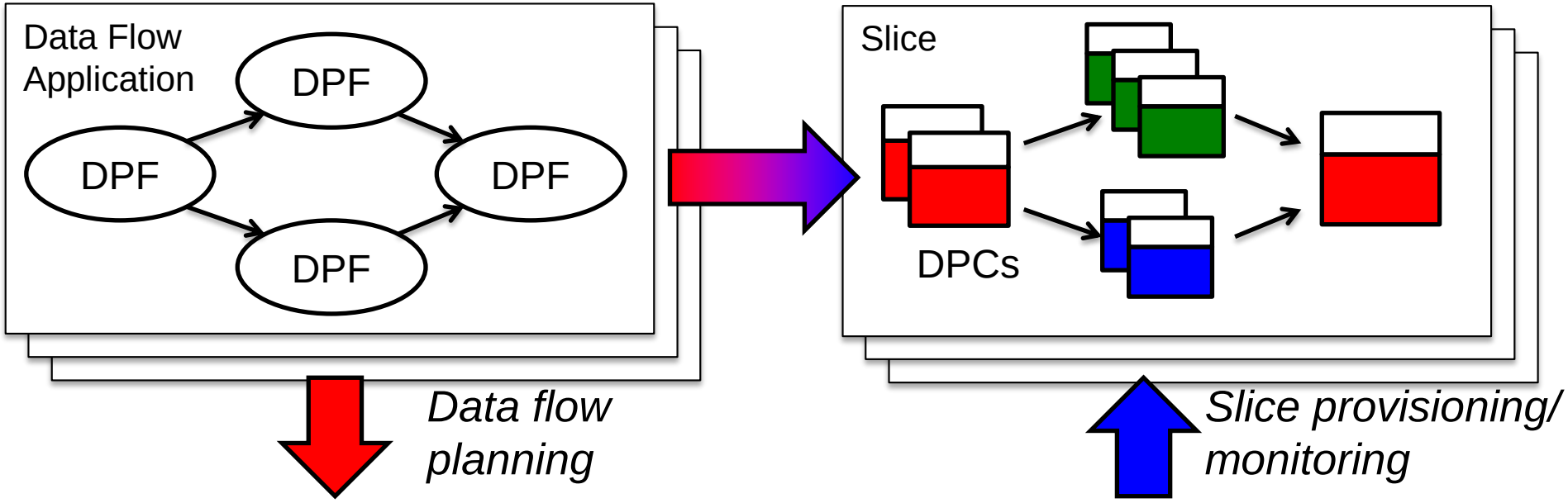
Wavelength Bank and Silicon Photonics

- AIST has been investigating photonics technology for future data centers.
- Wavelength Bank (WB: a centralized generator of wavelengths for DWDM) – light source
- Silicon photonics are used for light wave processing at a node, enabling hybrid implementation with electronics
- DWDM signals can be switched in one bundle by fiber cross-connect switches

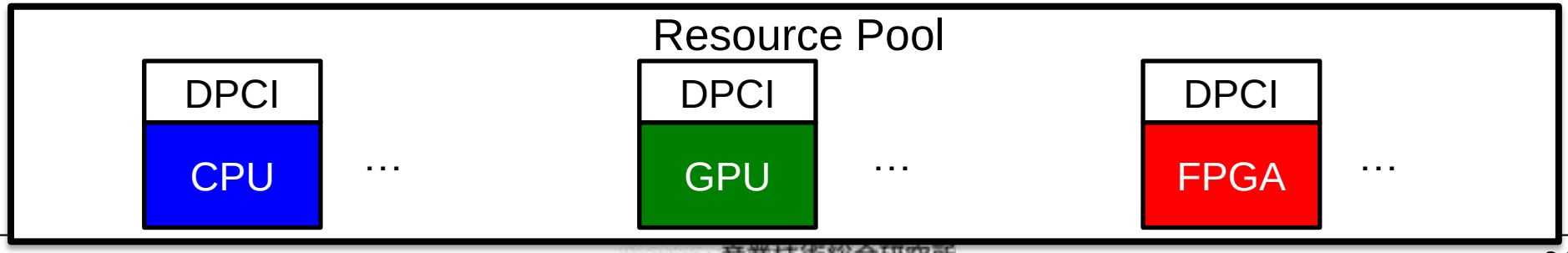


Data Flow OS

- Logical planning and physical mapping of data flow
- Resource monitor and management



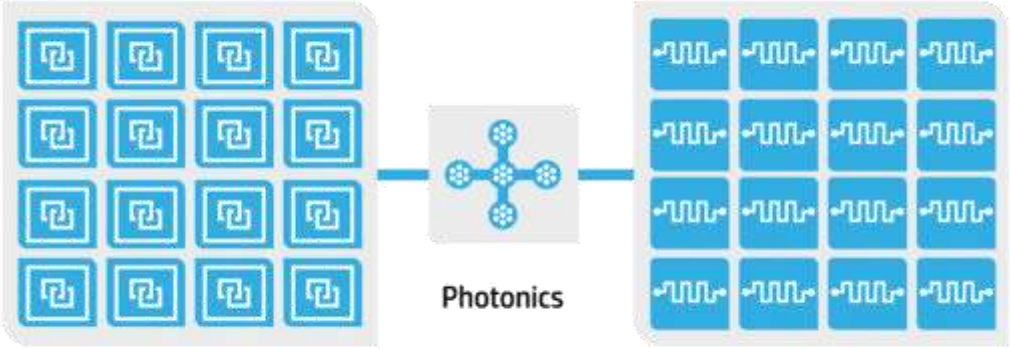
Data Flow OS



Related Work

HP The Machine

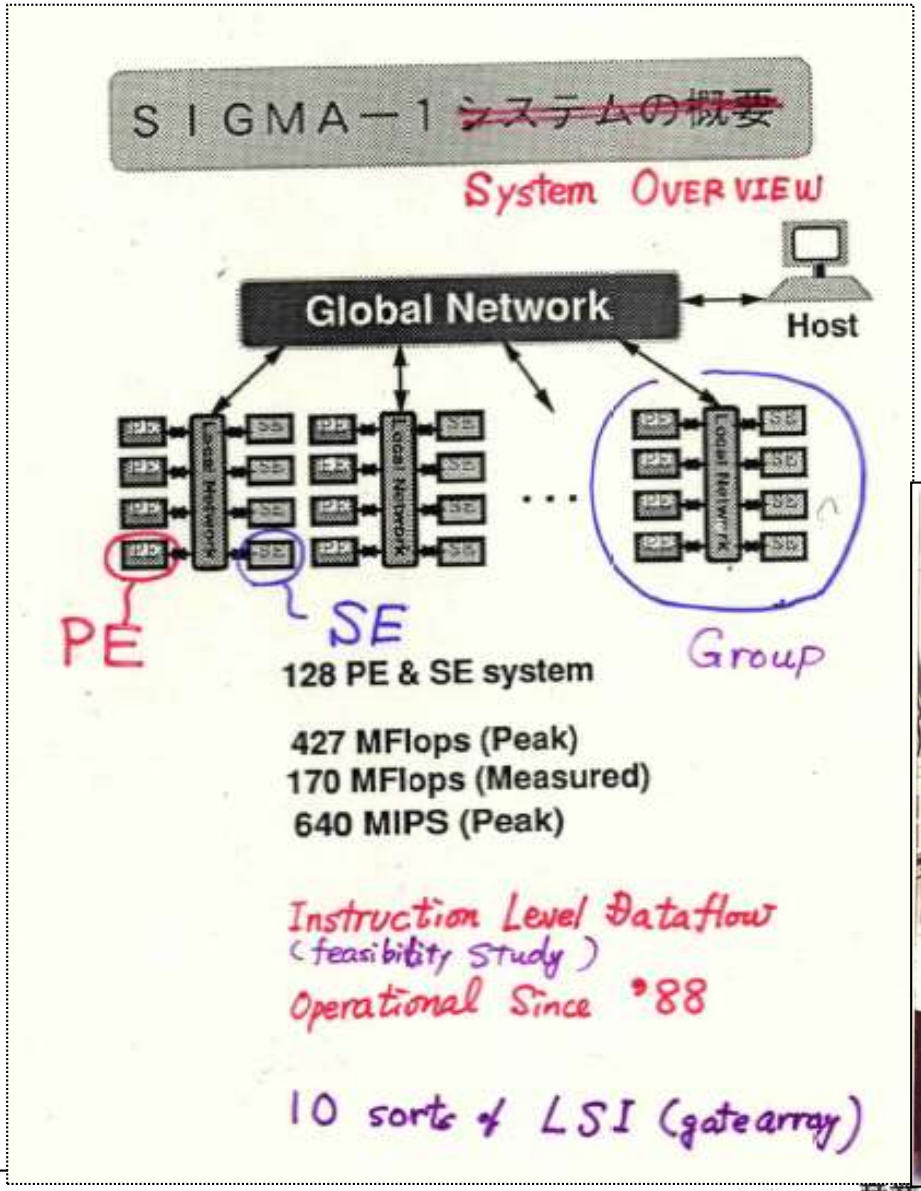
- Announced on June 11
- Combines many kinds of specialized cores to tackle different tasks as best as possible
- Cores are connected with non-volatile memory components using photonic network
- New OS will be developed by simplifying Linux and reducing overhead



Google Cloud Dataflow

- Announced on June 25
- New services that help developers build and optimize data pipelines
- A fully managed service for creating data pipelines that ingest, transform and analyze data in both batch and streaming modes
- Cloud Dataflow is a successor to MapReduce

Dataflow Supercomputer SIGMA-1 for scientific computation in 1988



- It worked!! With 128PE/SE
  - Peak 427MFlops
  - Measured 170MFlops
  - Cray X-MP
  - Processor, Network, OS, Language, Application
  - Dataflow-C

Summary

- The Big Data is drawing public attention.
- Big Data Example
  - Satellite Imagery Data is useful for planning disaster mitigation
  - Changing the game will produce more applications
- Workflow/Schematic Editor is useful for handling streaming data
  - Image processing, change detection
- Real time big data processing platform
  - Data movement is a key challenge
  - Photo networking – with less overhead
  - Dataflow revisited