Experimental Big Data Platform: Developments and Applications to Environmental Monitoring

Whey-Fone Tsai, National Center for High-performance Computing
Email: wftsai@nchc.narl.org.tw

NARLabs
National Applied Research Laboratories
National Center for High-performance Computing

Outline

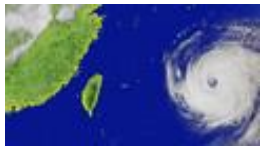
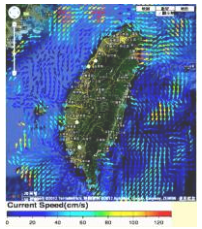
- **Ongoing big data applications in disaster management and environmental monitoring undertaking at NCHC.**
- **Experimental Big Data Platform for Environmental Monitoring**
  - ✓ **Global Climate Data Visualization**
  - ✓ **Air Quality Assessment and Response**
- **Concluding Remarks**

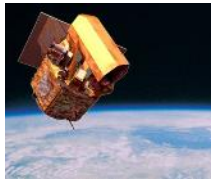
NCHC Big Data in Disaster Management and Environmental Monitoring(1)

Volume , Variety, Velocity, Veracity

Peta-Scale Earth Science Database

- **Integration/Archives of NARL's observation science data**
  - ✓ Ocean Observation database
  - ✓ Satellite: FORMOSAT-2and 5 image
  - ✓ Typhoon & Flood database
  - ✓ Earthquake Monitoring data




- **QA/QC implementation**
- **Homogenized database SQL**
- **Data provided for applications**

EU Fish4Knowledge

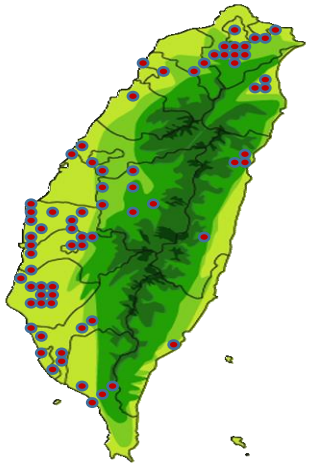
- **Long-term monitor of fish assembly in coral reef at Southern Taiwan**
  - ✓ Data: 8 undersea cameras for image recording in 3 years
  - ✓ Video streaming using Storm
- **Application for fish specie recognition**


- **Statistic analysis information**

Flood and Inundation Monitoring

- **>1100 CCTV (Water Resource Bureau)**
- **Image pattern recognition**
  - ✓ Automatic Flood and inundation detection

- **Big Data Analytics** (Bayes Theorem)
  - ✓ Predicting timing of CCTV malfunction.

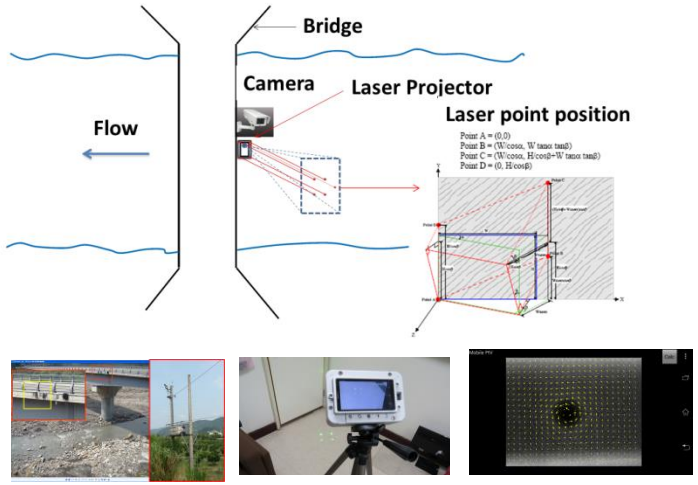


NCHC Big Data in Disaster Management and Environmental Monitoring(2)

Volume , Variety, Velocity, Veracity

River Flow Velocity Analysis

- **Using Laser-Positioning PIV** (Particle Image Velocimetry) monitoring system to measure movement of water surface particles
- **Data analysis of velocity** on NCHC Braavos Hadoop big data platform (302 nodes; 2PB)



- **1TB data 216 cores; speedup 221**; Computing time 1hr; 10days-single node.

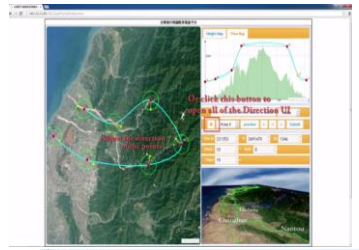
Disaster Management Information Platform for Research and Application

- Integrating 31 government open data
- 4 value-added service: data, model, display and management
- 6 scenario-based applications: flood, landslide, climate change, Earthquake, water resource & drought, disaster information

- **Web service** http://dmip.tw

i-Flyover GIS navigation

- **Data: Taiwan GIS database** (satellite and disaster event image)
- **Enabling Large GIS data visualization**
  - ✓ Navigate 3D geographic terrain from any direction, angle and height
  - ✓ Produce video on line
- **Applications:**
  - ✓ Disaster management
  - ✓ National spatial development plan

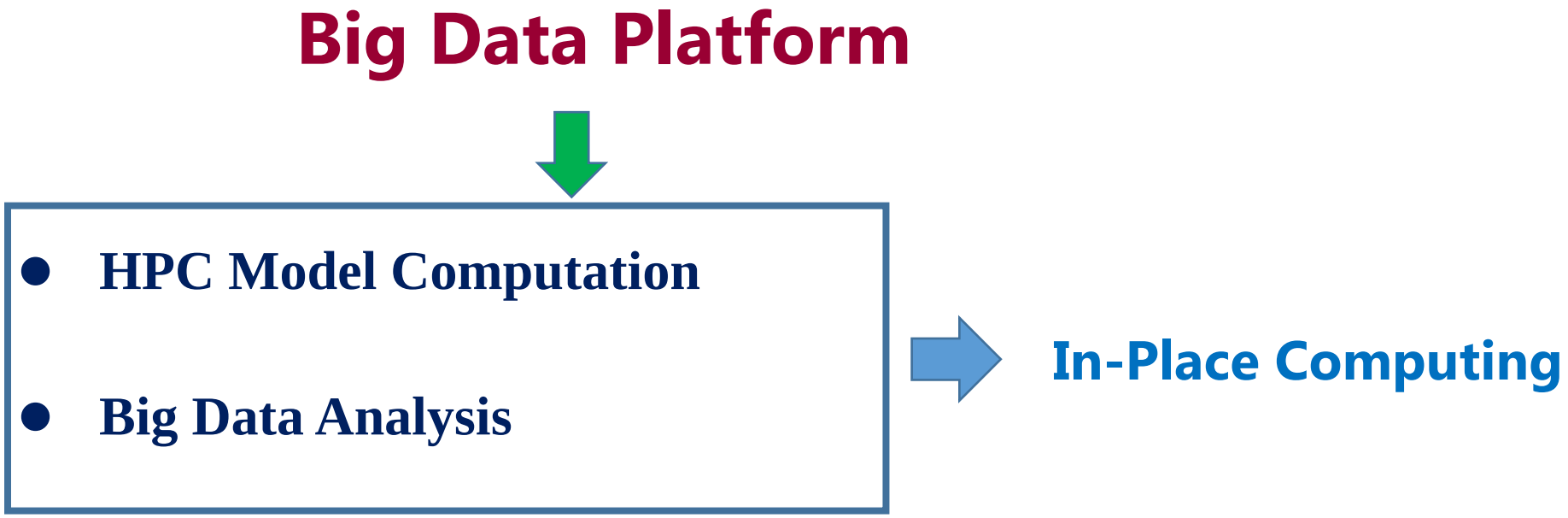


Formosat-2 Satellite

- **Web service** http://iflyover.tw

Environmental Monitoring: Features of Data

- **Observation** SENSOR DATA
- **Forecast** HPC_COMPUTATION DATA
- **Display**: Visualization of Simulation Data BIG DATA Visualization/Analysis



Collaborators

NCHC

NARLabs

Jen-Gaw Lee
Whey-Fone Tsai
Lung-Cheng Lee
Ching-Yao Lin
Hsi-Ching Lin
T.-M. Chuang

國立中興大學
National Chung Hsing University

Ben-Jei Tsuang, Environmental Engineering
I-En Liao, Information Science and Technology
Kai Chen Ku, Environmental Engineering
Shinyii Liu, Environmental Engineering
Wen-How Cheng, Information Science and Technology

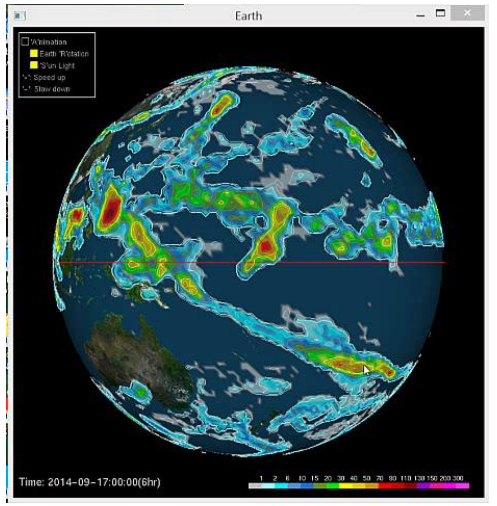
中央研究院
ACADEMIA SINICA

Chia-Ying Tu
Environmental Change Research

HPC & Big Data for Global Climate Model Simulation

- **HPC model simulation creates huge data**
  - ✓ EHTW global climate model (B. J. Tsuang) for 45-days Forecast, Grid size 125KM
  - ✓ NCHC ALPS 96 nodes, Computational time 7 hours; Output data around 20 climate variables
  - ✓ Long-term forecast: accumulation of data size > 200TB
- **Data visualization is crucial**
- **Look deep insight into climate huge data**
  - ✓ Feature Detection from Visual Animation
  - ✓ Assessment for pre-warning and response

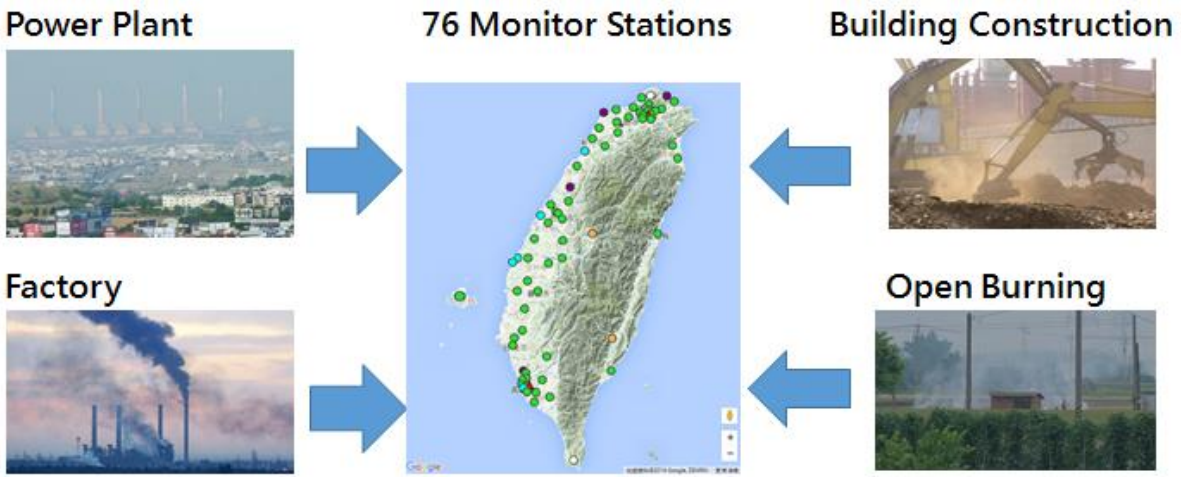
1. Precipitation and Sea level pressure for Asia
2. Precipitation and Soil Wetness for Asia
3. Vertically integrated water vapor and Soil Wetness for Asia
4. Vertically integrated water vapor and Precipitation for Asia
5. Sea level height and 50m current for South China Sea
6. Surface runoff into ocean and salinity for South China Sea
7. Sea Surface Temperature and Precipitation for Pacific Northwest
8. 100m current and Sea Surface Temperature for Pacific Northwest
9. Sea ice, Sea level height and 50m current for North Pacific
10. Sea ice, Sea level height and 50m current for West North
11. Sea ice, Sea level height and 50m current for South North
12. Sea ice, Sea level height and 50m current for Indian
13. 2m temperature
14. Precipitation
15. Sea level pressure and Precipitation
16. 100m current and Sea Surface Temperature
17. 100m current and Sea Surface Temperature
18. 10m wind speed
19. 200hPa Geopotential Height
20. 500hPa Geopotential Height



Air Quality Monitoring

- **General Pollutant Standards Index (PSI):** SO2, CO, O3, PM10, O3, NO2
- **Fine Particulate Matter: $PM_{2.5}$**
- **Air Quality Forecast/Response Assessment**
  - ✓ Wind Field Data
  - ✓ Forward trajectory simulation: point source's spatial influence
  - ✓ **Backward trajectory simulation**: contribution of point source pollutants on specific monitor station
  - ✓ **Response**: If forecast PM 2.5 is over critical value (purple color)
    - –Reduction of emission in point source
    - –Reduce the production of power plant and factory
    - –Response simulation can be made in advance

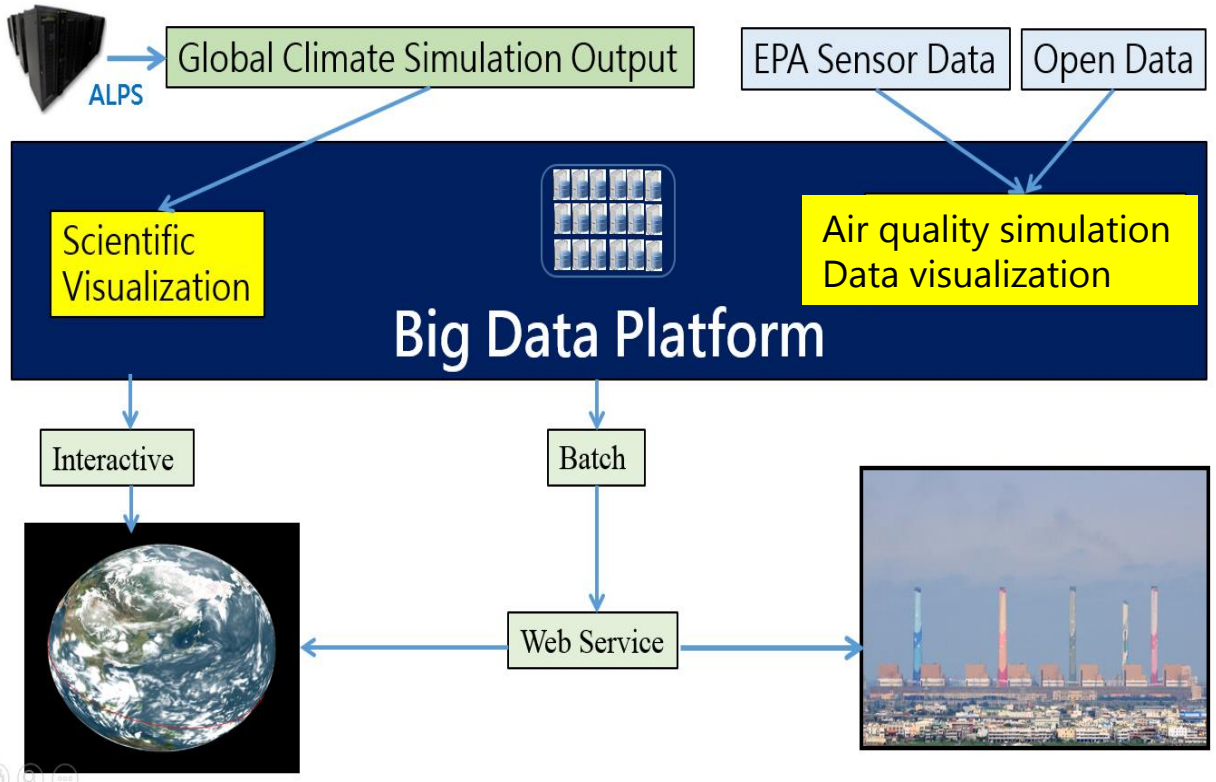
Index	1	2	3	4	5	6	7	8	9	10
Air Pollution Banding	Low	Low	Low	Moderate	Moderate	Moderate	High	High	High	Very High
$PM_{2.5}$ concentration ($\mu g/m^3$)	0-11	12-23	24-35	36-41	42-47	48-53	54-58	59-64	65-70	≧71



Taichung Thermal Power Plant is the world's largest coal-fired power plants

Experimental Big Data Platform: Conceptual Design

- **Structured Data Format**
  - ✓ Climate and air quality model output data are of spatial mesh data
- **In-Place Computing**
  - ✓ Currently, Hadoop streaming is still tedious to implement large-scale numerical code
  - ✓ Using the structured database SQL; and
  - ✓ Master-Slave distributed computation and data processing framework
- **Link to NCHC ALPS supercomputer**
  - ✓ If computational demands exceeds platform capacity



System Hardware and Software

Hardware

ALPS Acer AR585 F1 Cluster
-177 TFLOPS
-25,600 compute cores

Supercomputer

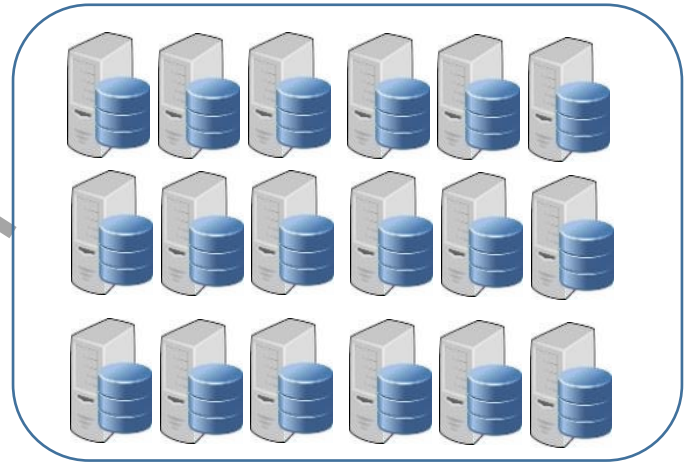
10GE

Storage Total : 160 TB
NAS Server TS-EC2480U-RP (24bay) - 96TB
QNAP REXP-1600U-RP (16bay) – 64TB

1 GE

Master Node

18X1 GE



18 Slave Nodes
ACER Veriton M2611G
CPU:Intel i7-3770@3.40GHz(4Coresx2Threads)
RAM:16GB
HDD: 4TB

Software

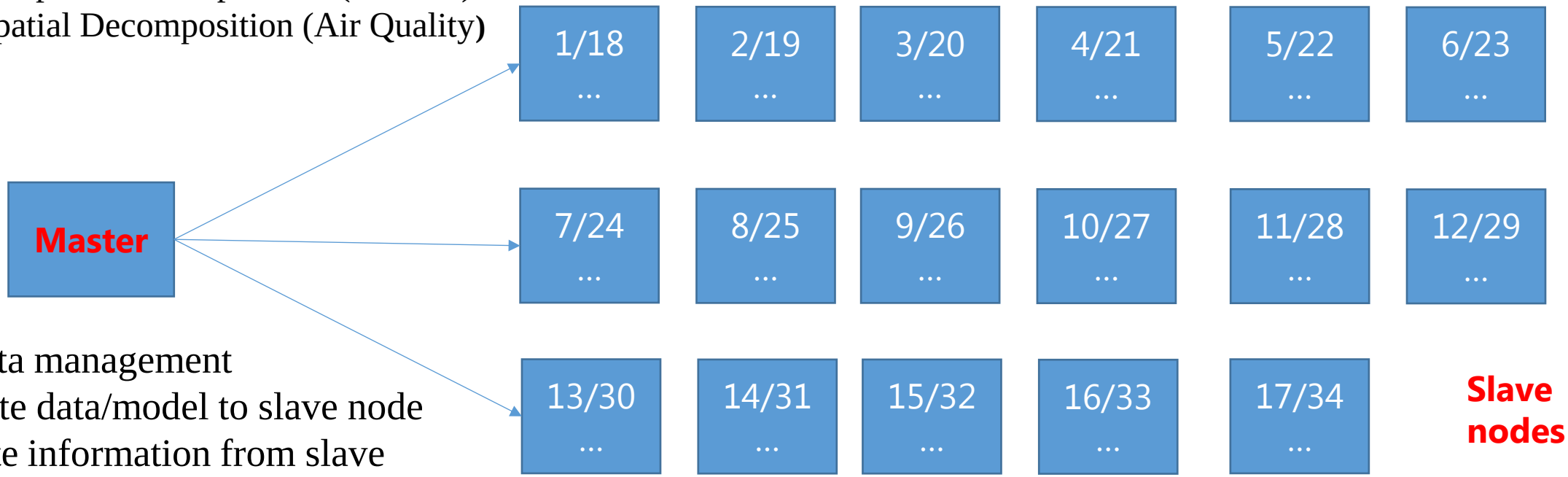
openSUSE

MariaDB

Linux + Mysql

Data Dispatch

- **Data Dispatch from Master Node to Slave Nodes**
  - ✓ Temporal Decomposition (Climate)
  - ✓ Spatial Decomposition (Air Quality)



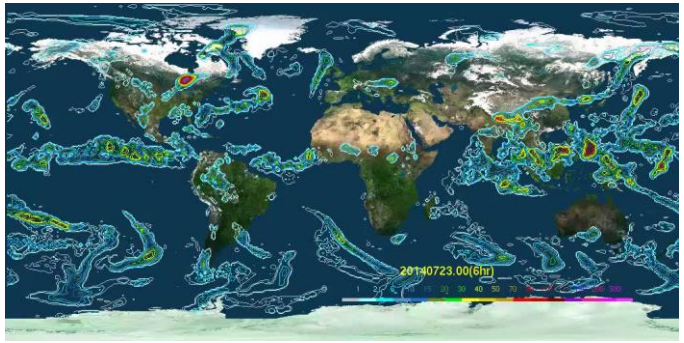
- Metadata management
- Replicate data/model to slave node
- Integrate information from slave nodes
- MPI

- Using DB to query computation or analysis data
- Independent data processing (no communication) among slave nodes
- Distributed processing jobs following the data node sequence

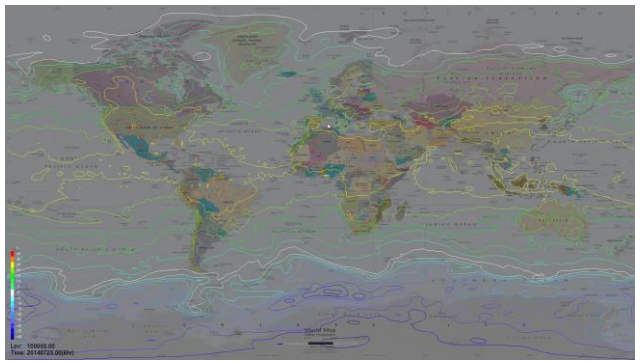
Global Climate Data Visual Animation

- **NCHC EHTW global climate model**
  - ✓**45**-days Forecast data: Precipitation, Temperature….
- **HiRAM** High Resolution Atmospheric Model
  - ✓ Outgoing Longwave Radiation (OLR) data
- Visualization: C/C++ and OpenGL
  - ✓ Two different display setups: 2D Map and 3D sphere background

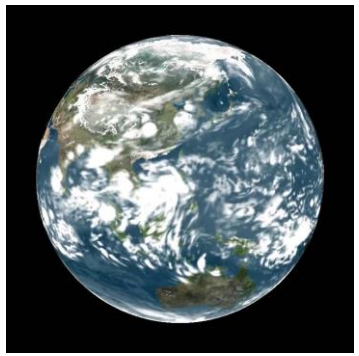
Data		OLR Simulation	Precipitation Simulation
Dimension		1536 x 768	320 x 160
Time steps		248	181
Sequential Computation		~670 sec	~13.2 sec
Distributed System	Nodes	17	17
	Computation	~58 sec	~2.3 sec
	Speedup	11.55	5.74

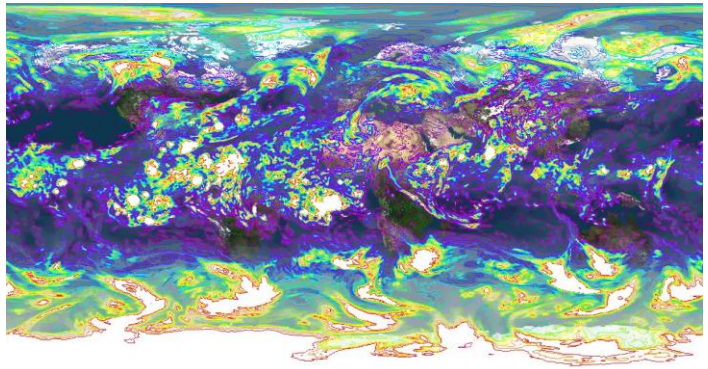


Precipitation



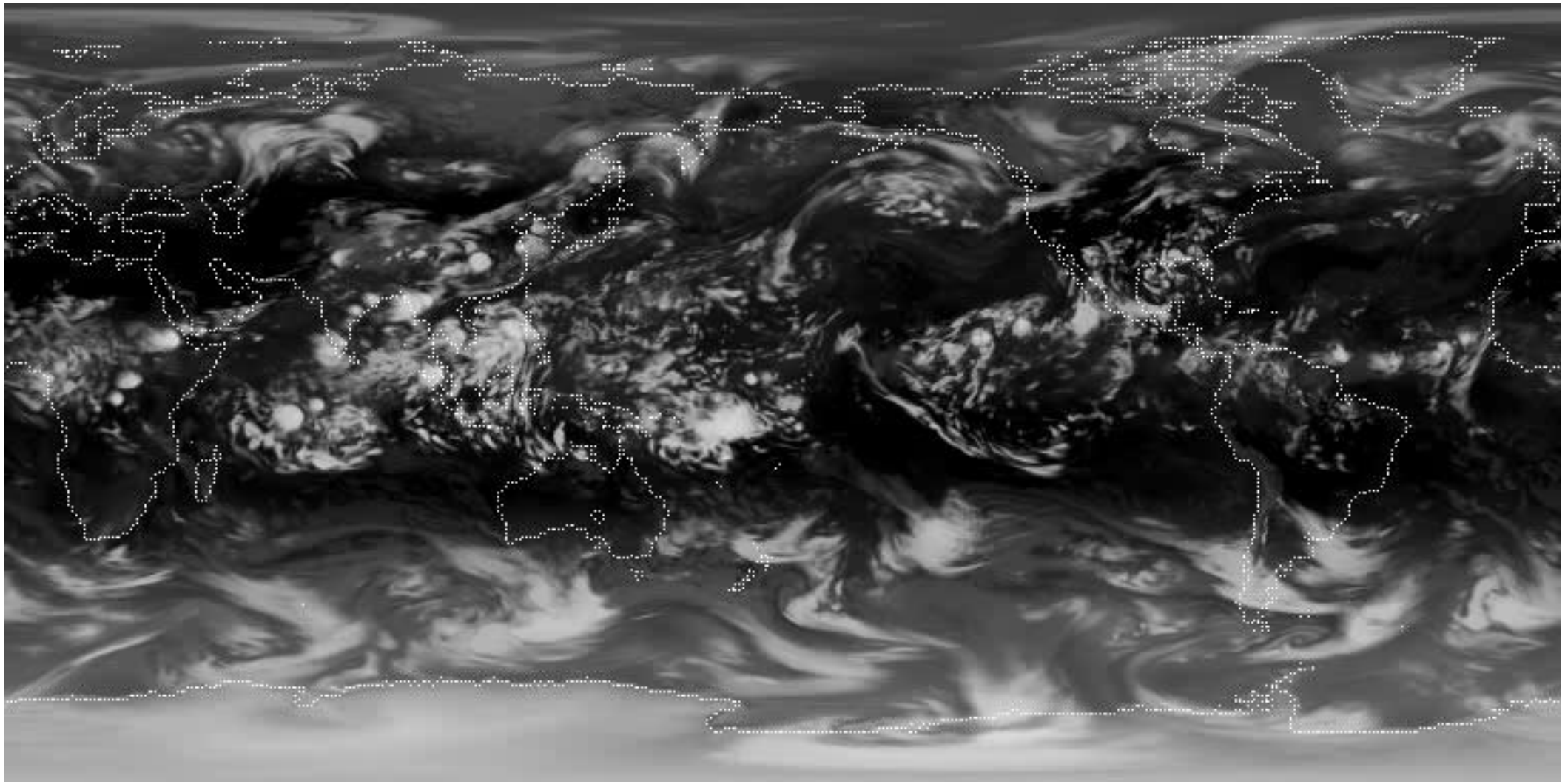
Temperature





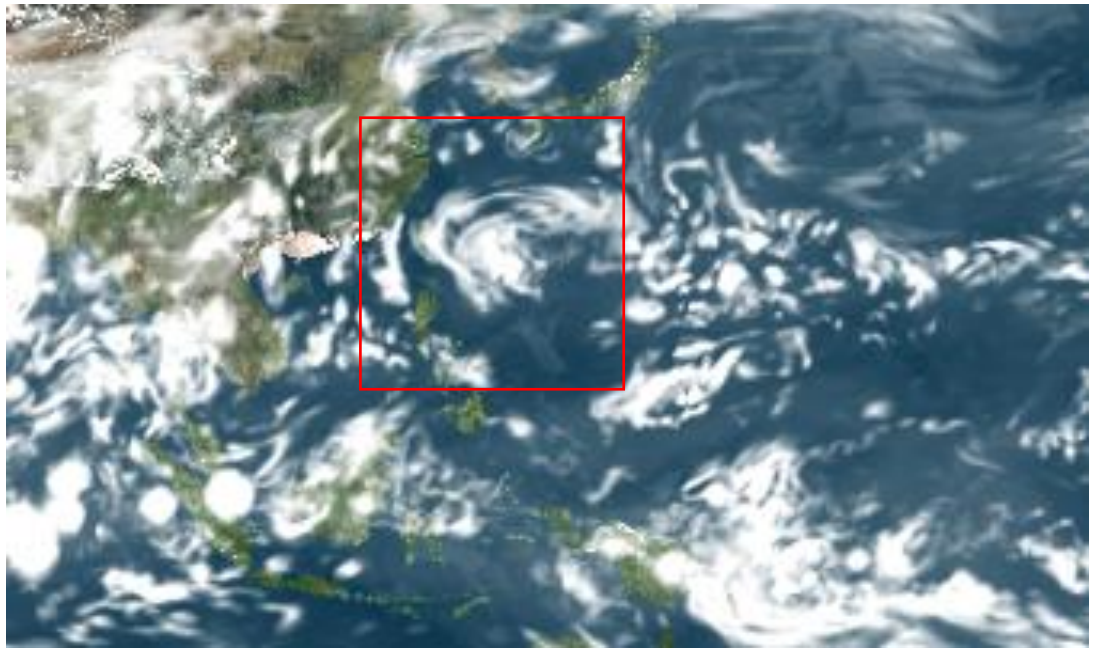
OLR (Index of Sunlight)

Feature Detection from Visual Animation



OLR simulation data

Feature Detection from Visual Animation



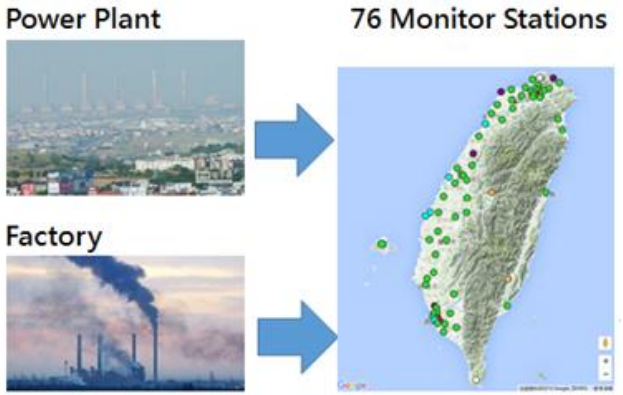
Formation of Typhoon eddies are detected through image pattern recognition

Downtown Hsinchu City snowed on Jan 23, 2016

Long-term global climate simulation may be possible to predict the something impossible before.

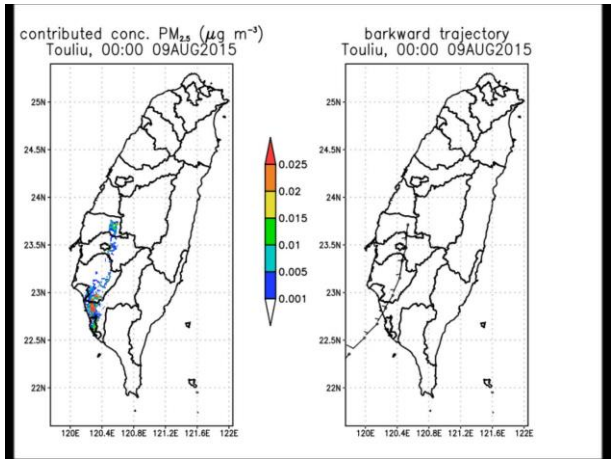
Air Quality Forecast and Response

- **Air Quality Forecast: $PM_{2.5}$**
  - ✓ Wind Field Data (EPA 2007)
  - ✓ **Backward trajectory simulation**: contribution of point source pollutants on specific monitor station. GTx air quality model (B.J. Tusang) is used.
- **Response**: PM 2.5 exceeds critical value
  - –Reduction of emission in point source
  - –Reduce the production of power plant and factory
  - –Response for re-simulation using updated information



2007 EPA Daily Wind Field for assessment

Backward trajectory simulation



Data Replication and Parallel Computation

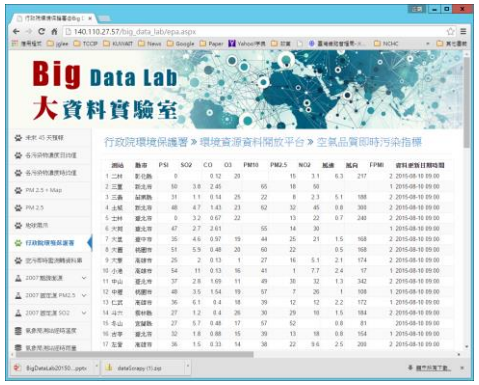
Government

EPA Open Data

Scheduler

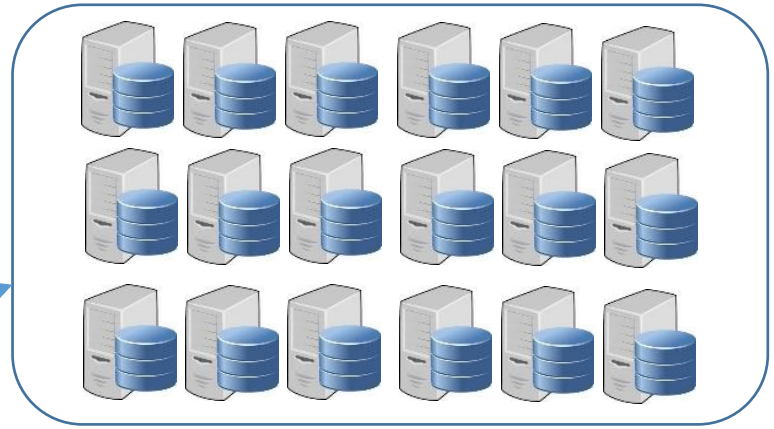
+Parser

DB



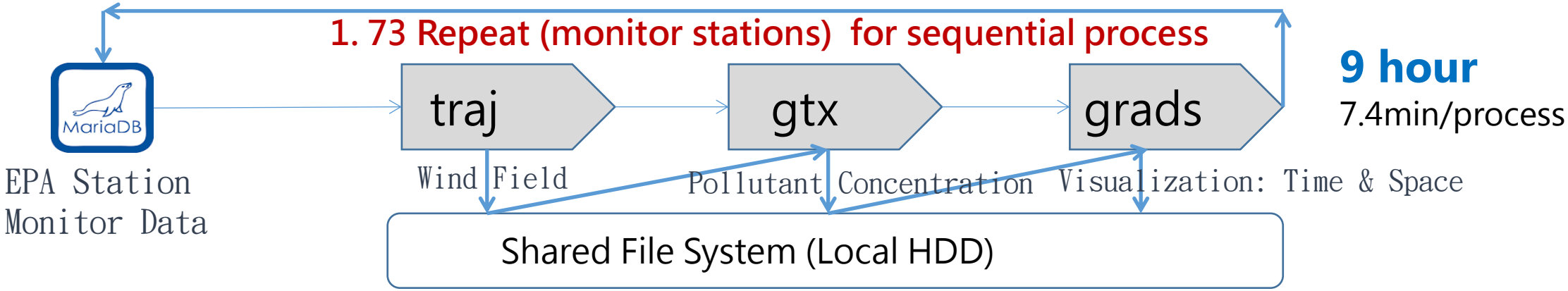
Master

Slave*18

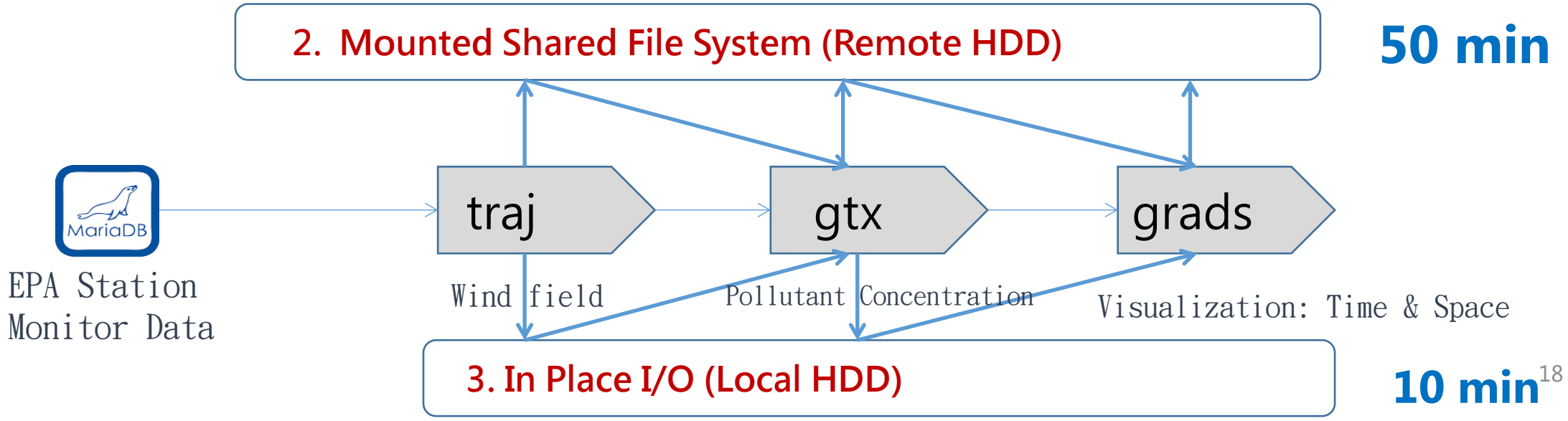


- **Each slave node has installed simulation code and necessary data from Master node**
- **Each slave node is in charge of model computation and visualization for fixed monitor stations**

Performance Benchmark

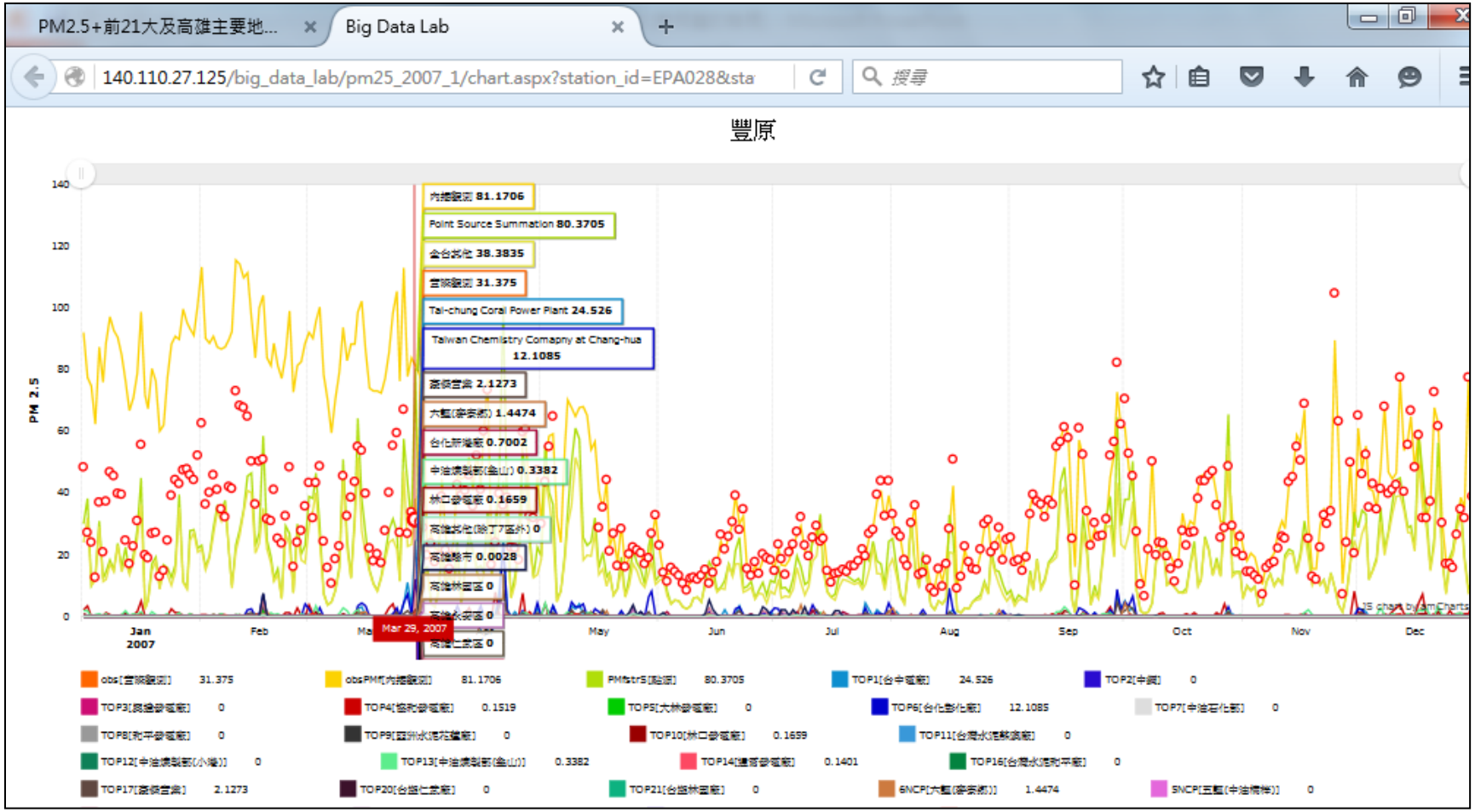


Parallel Computation : 73 cores, 17nodes

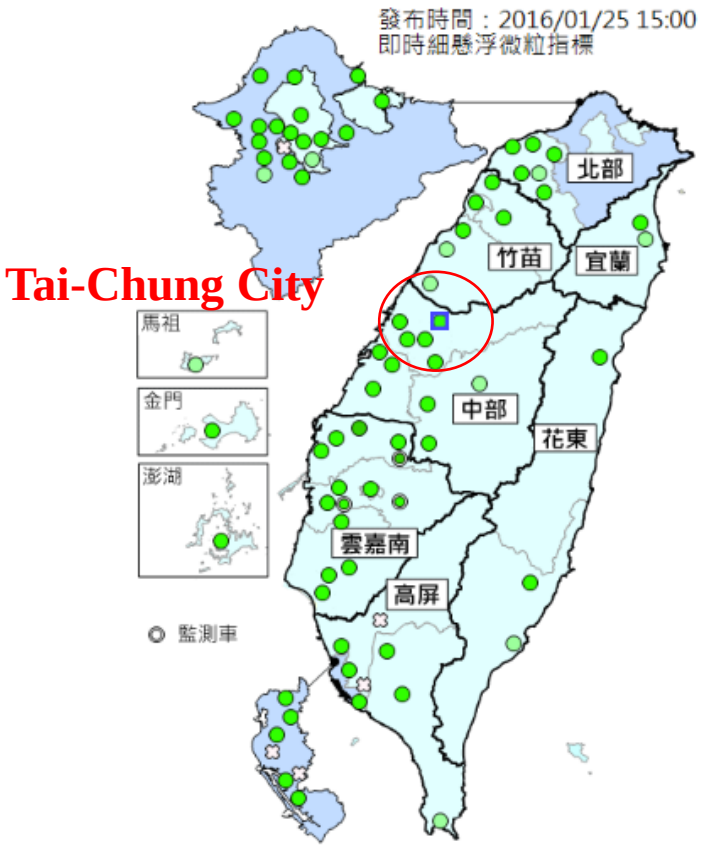


User-friendly, Web Service Based Big Data Presentation

- **Website covering** the heterogeneous observed and forecast tedious pollutant data in space and time
- **Supporting** for decision making and emergency response in near real time.
- **Website development**
  - ✓ ASP.Net 4.5 programming language hosted on Microsoft Windows server 2012 R2
  - ✓ MariaDB database
  - ✓ AMCHARTS JavaScript charting library draw
  - ✓ google map API



Case Study of Air Quality Forecast and Response



請點擊左方測站位置或
請選擇：中部 > 豐原 查詢

豐原測站

PSI 空氣污染指標		設備維護
PM_{10} 懸浮微粒 ($\mu g/m^3$)	移動平均值	
	小時濃度值	29
O_3 臭氧 (ppb)	小時濃度值	43
$PM_{2.5}$ 細懸浮微粒指標		2 低
$PM_{2.5}$ 細懸浮微粒 ($\mu g/m^3$)	移動平均值	12
	小時濃度值	16

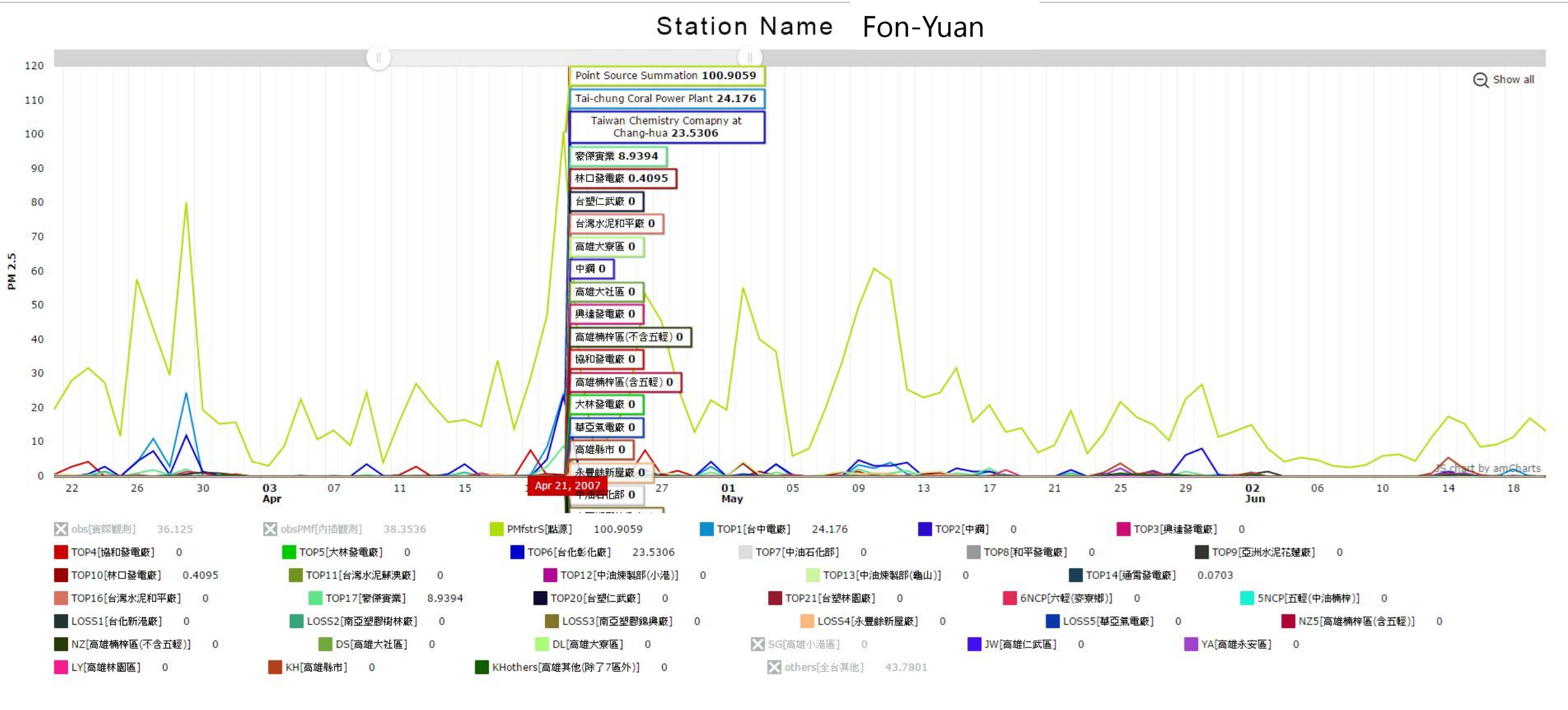
單位：1.$\mu g/m^3$，微克/立方公尺
2.ppb，十億分之一
：設備維護（測站例行維護、儀器異常維修、監測數據不足）

Tai-Chung thermal power plant

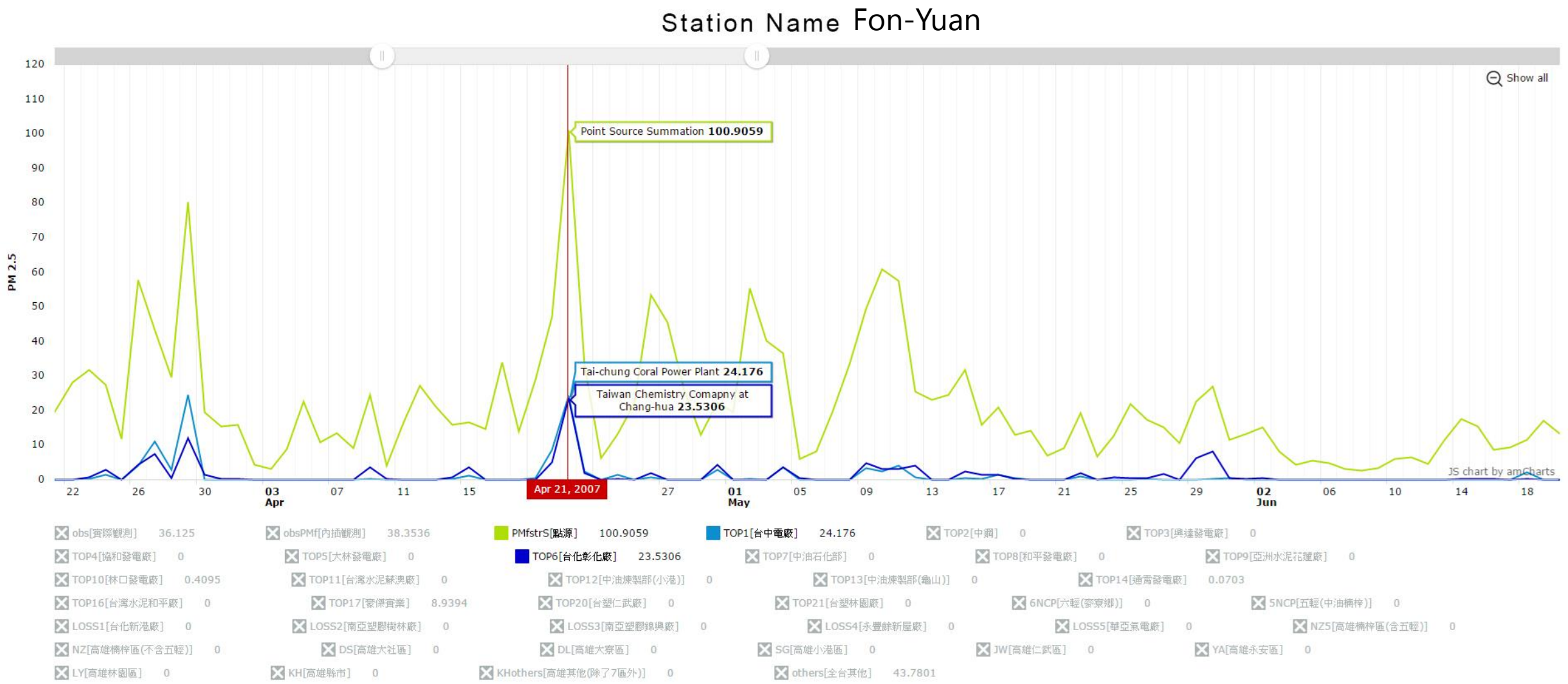
5 Chimneys and 10 power generators

Fen-Yuan Monitor Station

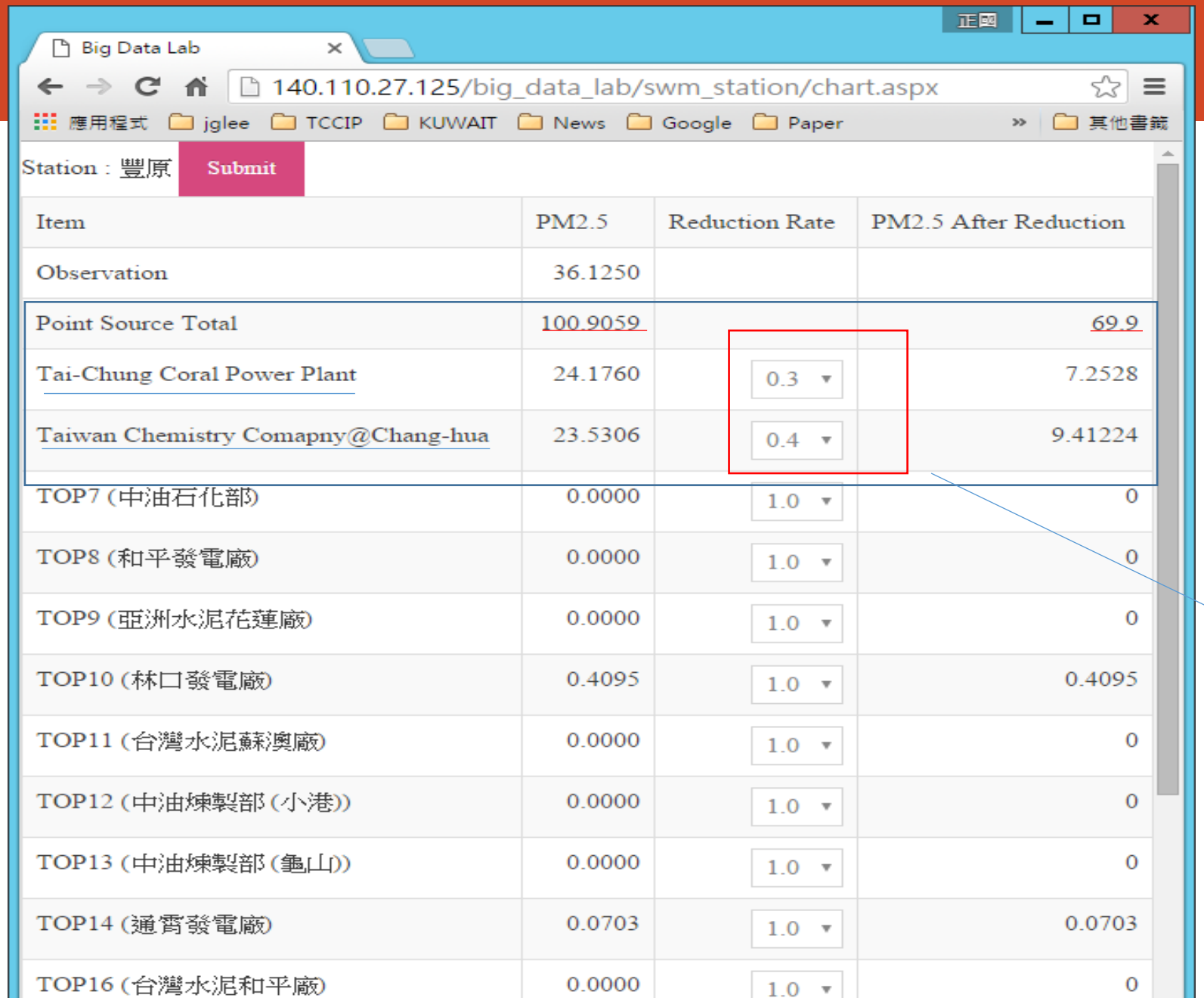
Response Assessment at Fon-Yuan Monitor Station



Response Assessment at Fon-Yuan Monitor Station



Reduction of Point Source Emission for Response

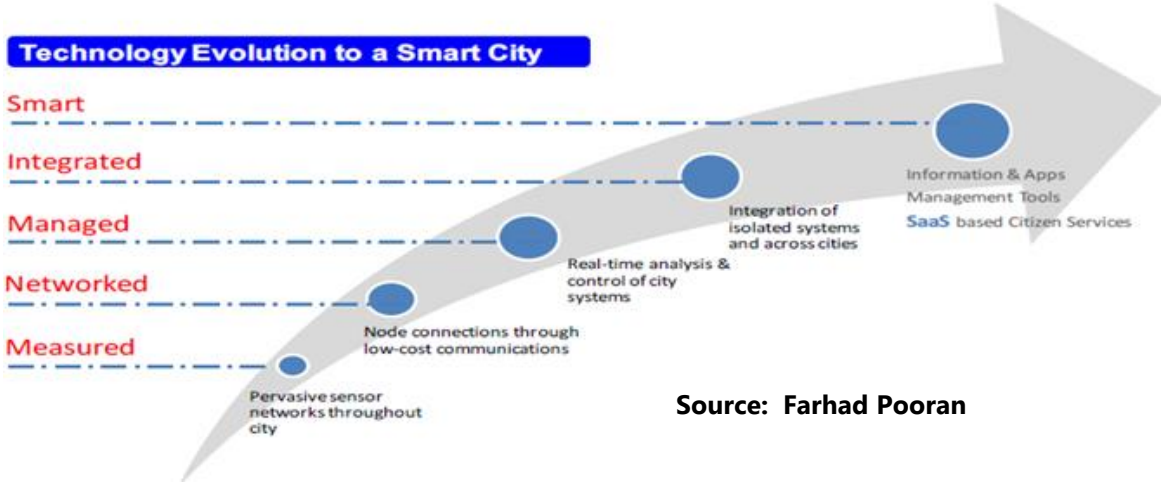


Item	PM2.5	Reduction Rate	PM2.5 After Reduction
Observation	36.1250		
Point Source Total	100.9059		69.9
Tai-Chung Coral Power Plant	24.1760	0.3	7.2528
Taiwan Chemistry Comapny@Chang-hua	23.5306	0.4	9.41224
TOP7 (中油石化部)	0.0000	1.0	0
TOP8 (和平發電廠)	0.0000	1.0	0
TOP9 (亞洲水泥花蓮廠)	0.0000	1.0	0
TOP10 (林口發電廠)	0.4095	1.0	0.4095
TOP11 (台灣水泥蘇澳廠)	0.0000	1.0	0
TOP12 (中油煉製部(小港))	0.0000	1.0	0
TOP13 (中油煉製部(龜山))	0.0000	1.0	0
TOP14 (通霄發電廠)	0.0703	1.0	0.0703
TOP16 (台灣水泥和平廠)	0.0000	1.0	0

Assessment results provided for government authority for action

Concluding Remarks

- **Scenario-based environmental monitoring applications covers the major ICT technologies: HPC, Network, Cloud, Big Data, IOT, CPS (Smart Big Data)**
- **Current developed experimental big data platform has been successfully applied for Taiwan Climate and Air Quality communities, and will further support for scenario-based environmental monitoring, such as air quality response associated with smart city management.**



Source: Farhad Pooran

Thank you