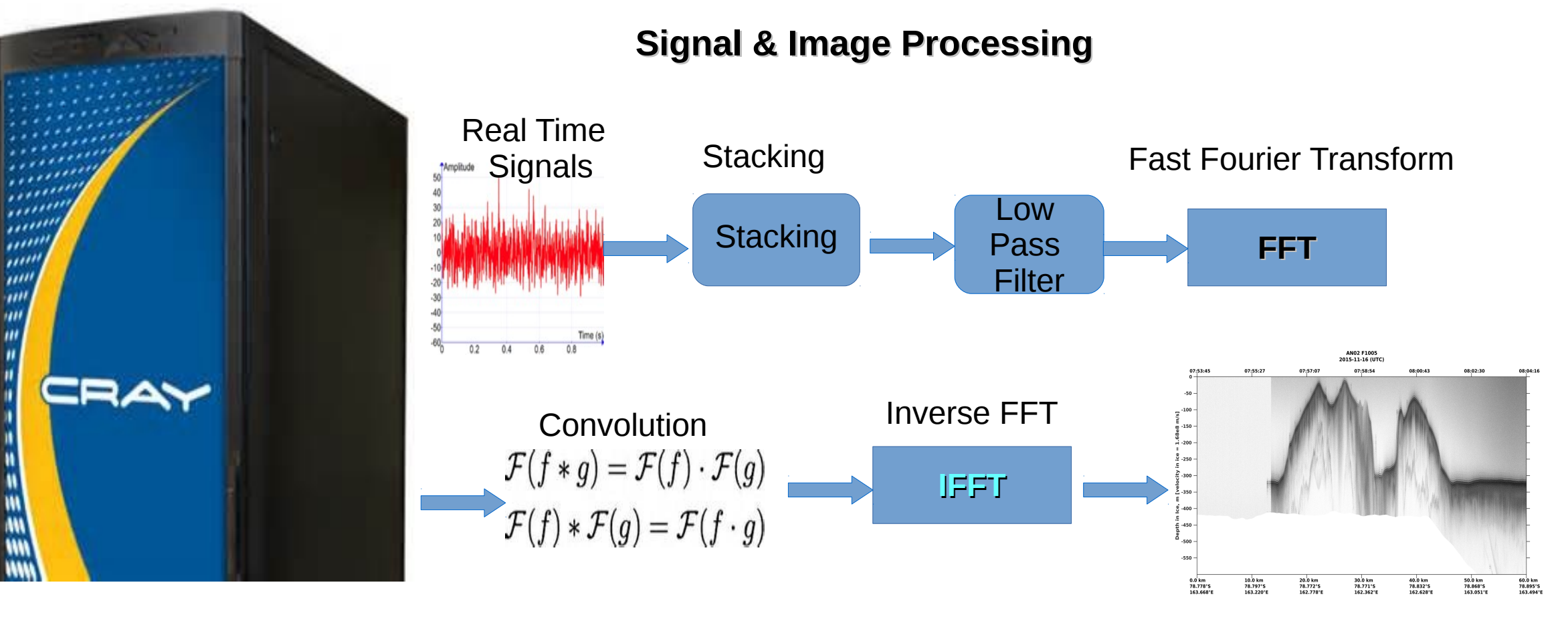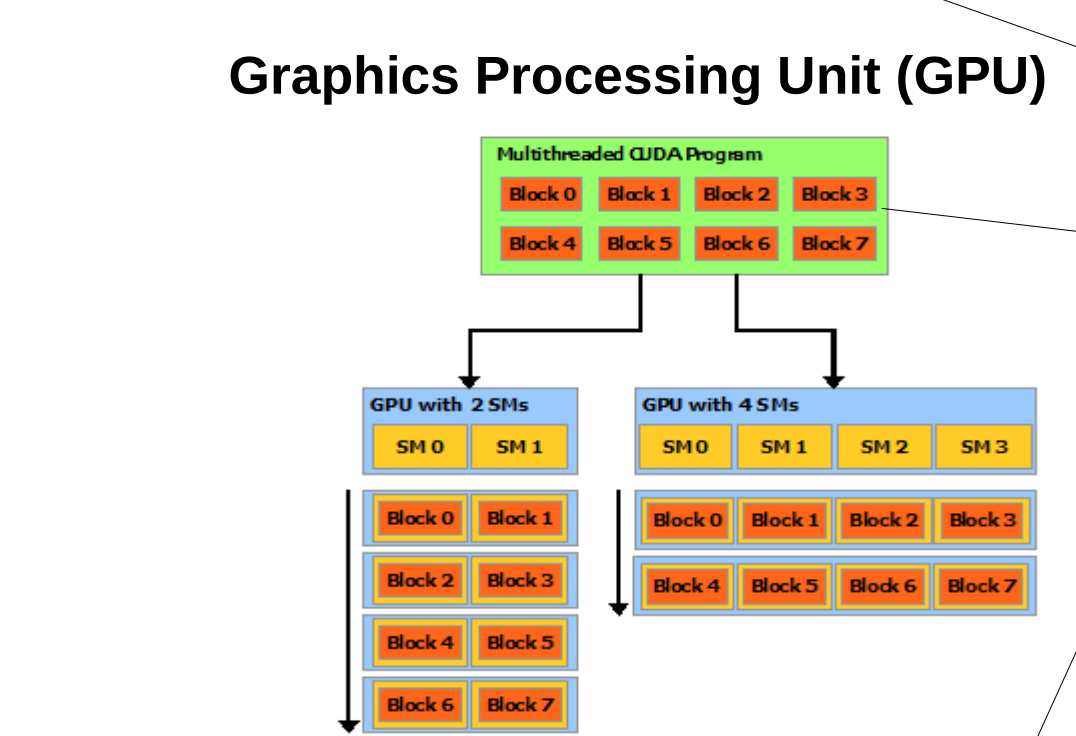# A Use Case of Big Data

LingLing Dong ldong@ldeo.columbia.edu   Nick Frearson nfre@ldeo.columbia.edu

Case 1. Improve Performance with Supercomputer Services
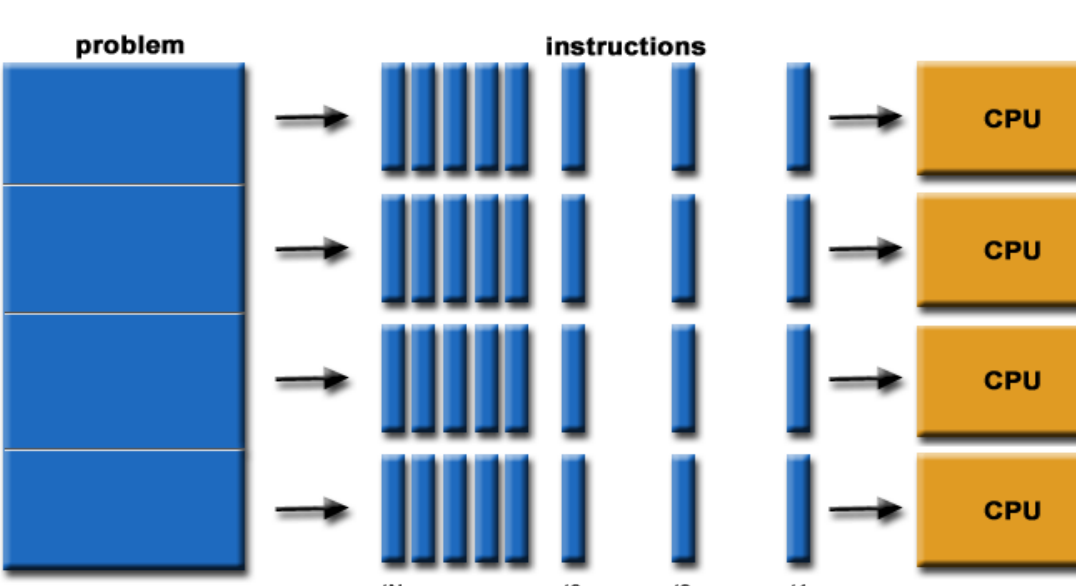
Supercomputer Saves Data Processing Time
Reduced the operation time from 48 hours to 3 hours

Signal & Image Processing

Real Time Signals → Stacking → Low Pass Filter → **FFT**

Convolution
$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$$
$$\mathcal{F}(f) * \mathcal{F}(g) = \mathcal{F}(f \cdot g)$$

Inverse FFT → **IFFT**

Multiple Processes

Supercomputers

Graphics Processing Unit (GPU)

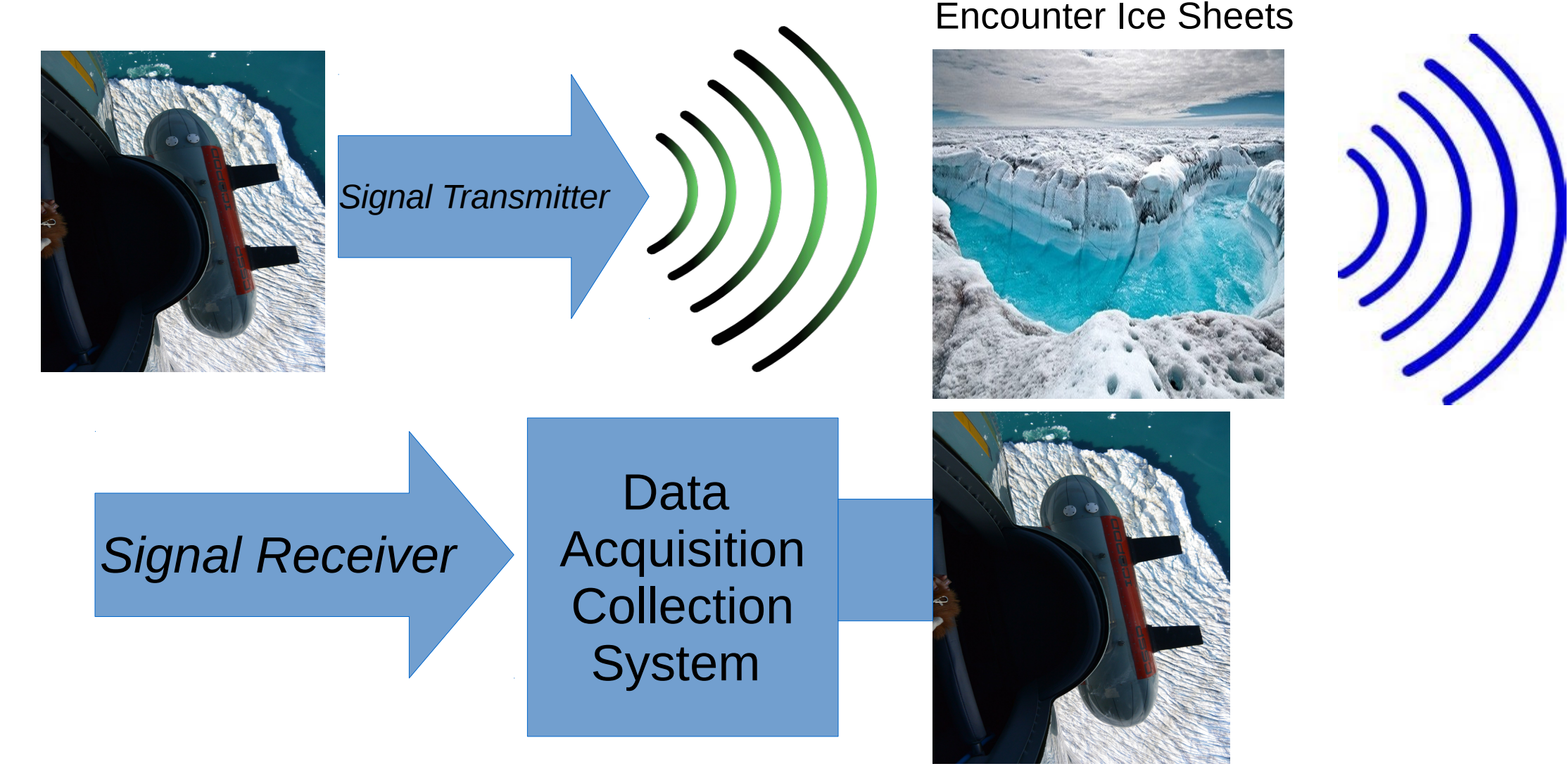#### Message Passing Interface (MPI)

The Hypothesis: Using super computer's high processors, GPU programming, and MPI to distribute the tasks in multiple CPUs, the data processing in a supercomputer would be much faster than in a single HPC.

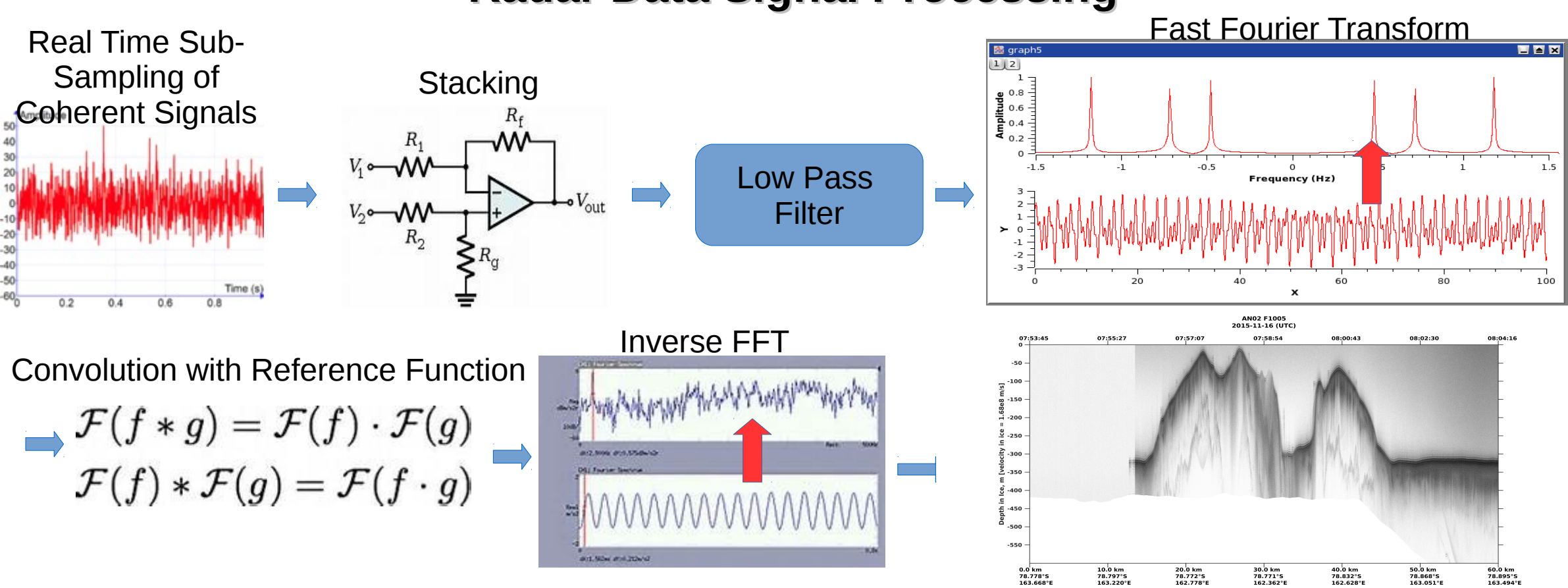**The Proof Result:** a data processing work that takes 48 hours can be done in 3 hours on a supercomputer.
Input/output Latency: Data IO may take 2 hours to up/download to supercomputers.

Modern science significantly depends on data and data technologies to quantitatively describe the objects under research. In our polar research, we employ a sophisticated set of instruments to study the ice-sheets. The data we collect and process comes to more than 100 TB a year across several physically distinct campaigns. This can be defined as big data. The technologies we apply through the phases of data collection, analysis, visualization, modeling, publication, and archiving invoke some new big-data machinery that we would like to share and discuss with other colleagues in different fields.
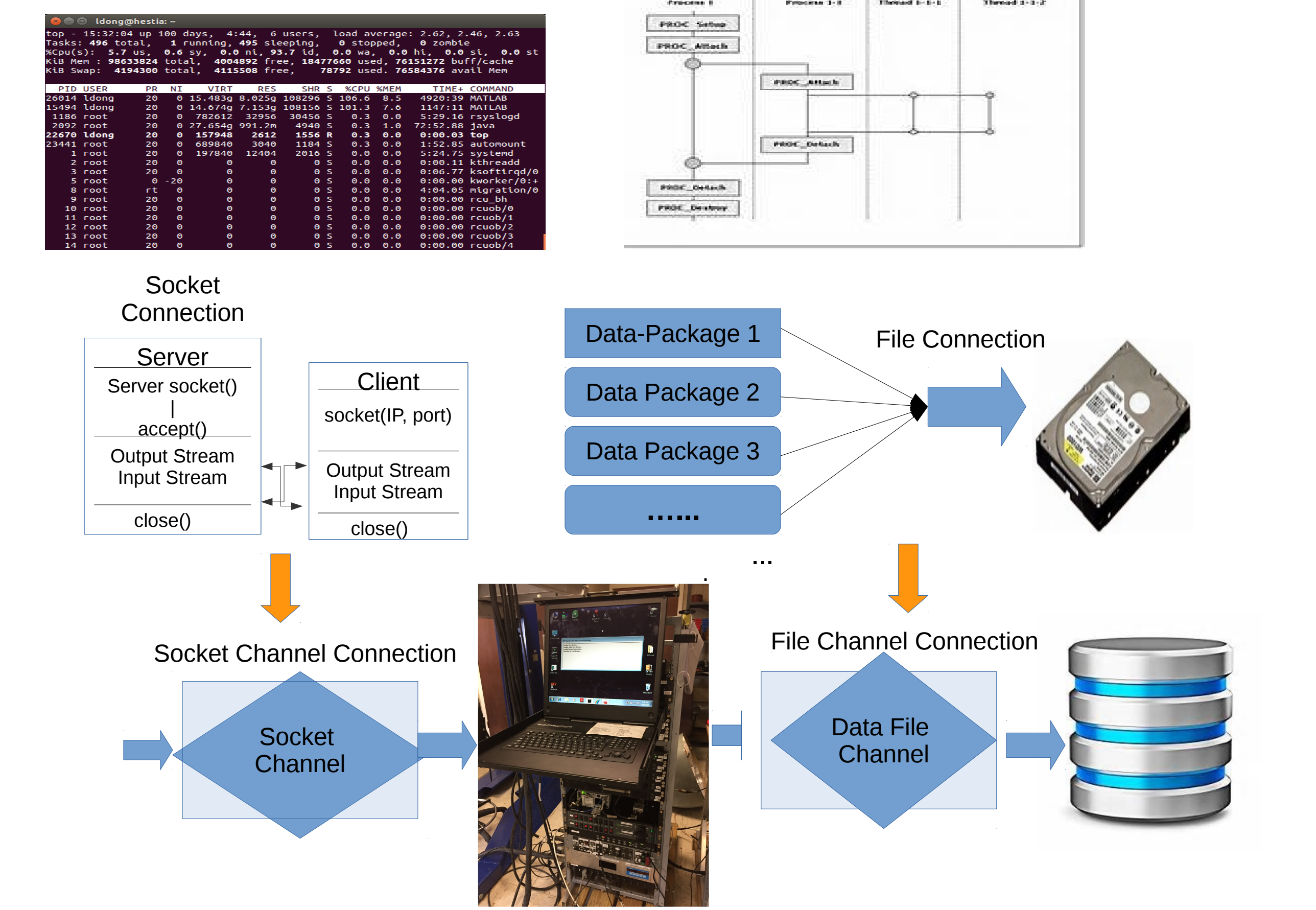
Field Radar Data Collection

Signal Transmitter

Encounter Ice Sheets

Signal Receiver → Data Acquisition Collection System

Radar Data Signal Processing

Real Time Sub-Sampling of Coherent Signals → Stacking → Low Pass Filter → Fast Fourier Transform

Convolution with Reference Function
$$\mathcal{F}(f * g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$$
$$\mathcal{F}(f) * \mathcal{F}(g) = \mathcal{F}(f \cdot g)$$

Inverse FFT

Case 2: Improve Performance in Big Data Operations

Resolved the Big Data Caused CPU Stress
Reduced CPU usage from 100% to 15%

Socket Connection

Server
Server socket()
accept()
Output Stream
Input Stream
close()

Client
socket(IP, port)
Output Stream
Input Stream
close()

Data-Package 1
Data Package 2
Data Package 3
......

File Connection

Socket Channel Connection → Socket Channel

File Channel Connection → Data File Channel

The Problem: During data acquisition, the data rate of the deep ice data radar can be as high as 86MB per second. This high volume of data caused data flow problems in the network. Using the ordinary socket method, the data is choked at the Internet port and stresses the data acquisition computer CPU.
The Resolution: Applied new programming technology of socket channel that opens a channel connection from the radar instrument to the computer, easing the data flow through the network. In addition to the new network connection, a file channel also opens to allow writing data to the data files on the fly.
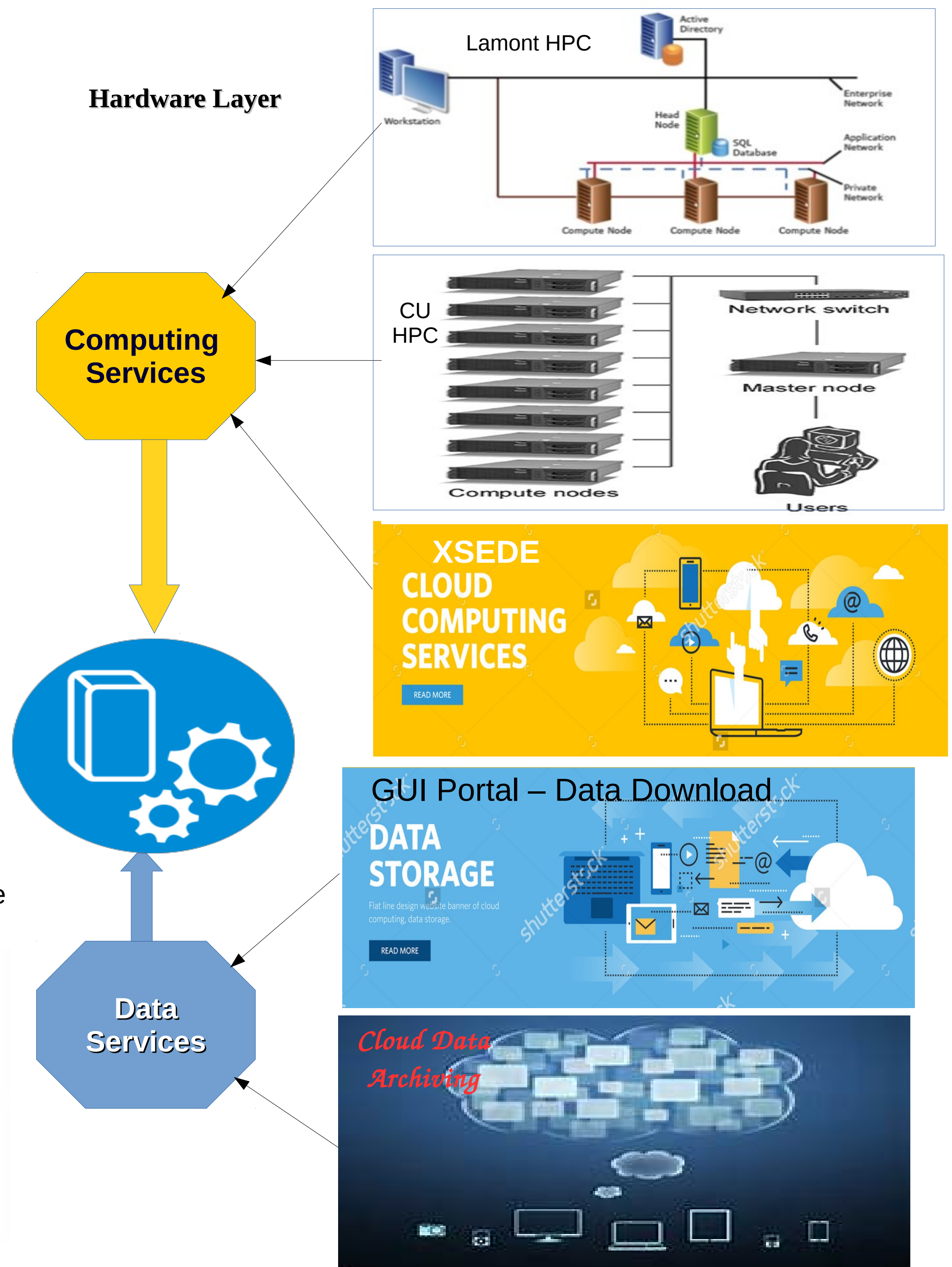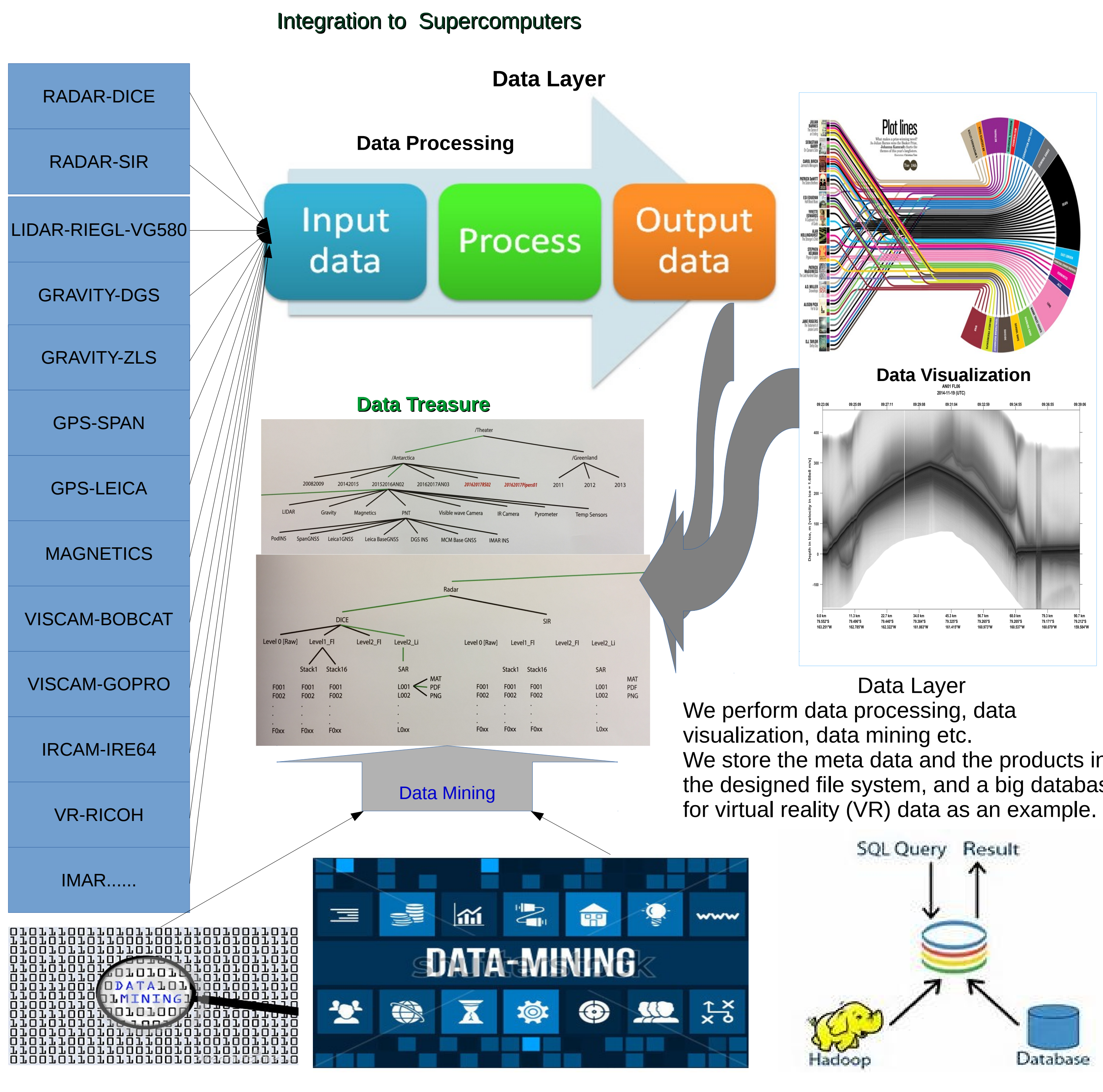The Result: The observation of the CPU usage goes down from 100% to 15%, and the performance remains good constantly.
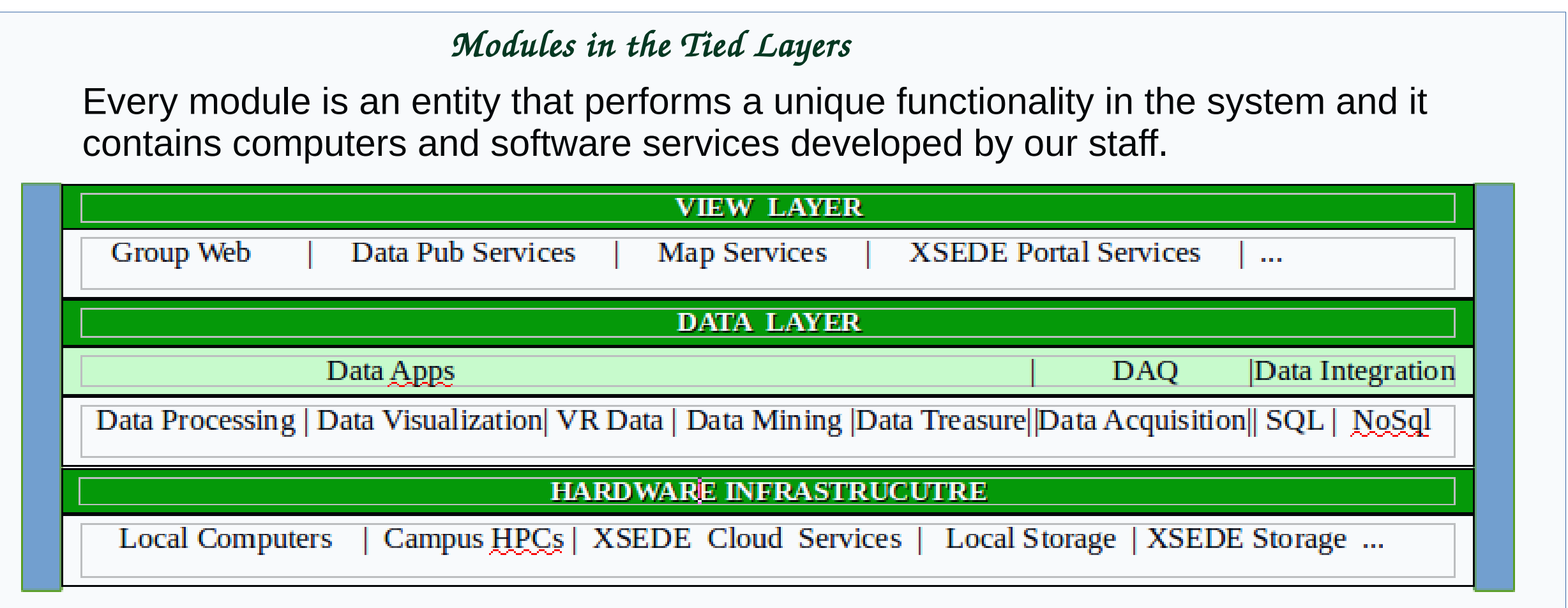
Integration to Supercomputers

Modules in the Tied Layers

Every module is an entity that performs a unique functionality in the system and it contains computers and software services developed by our staff.

VIEW LAYER			
Group Web	Data Pub Services	Map Services	XSEDE Portal Services	...

DATA LAYER		
Data Apps	DAQ	Data Integration
Data Processing	Data Visualization	VR Data	Data Mining	Data Treasure	Data Acquisition	SQL	NoSql		

HARDWARE INFRASTRUCTURE				
Local Computers	Campus HPCs	XSEDE Cloud Services	Local Storage	XSEDE Storage ...

View Layer

Map Server

Data Server

XSEDE Portal

The Polar Geophysics Group web server, wonder.ldeo.columbia.edu, will host services from the map servers, the data publication servers, and our science portal in the XSEDE supercomputer center.

Data Layer

RADAR-DICE
RADAR-SIR
LIDAR-RIEGL-VG580
GRAVITY-DGS
GRAVITY-ZLS
GPS-SPAN
GPS-LEICA
MAGNETICS
VISCAM-BOBCAT
VISCAM-GOPRO
IRCAM-IRE64
VR-RICOH
IMAR......

Data Processing

Input data → Process → Output data

Plot lines

Data Visualization

Data Treasure

Data Mining

DATA-MINING

SQL Query   Result

Hadoop   Database

Data Layer
We perform data processing, data visualization, data mining etc.
We store the meta data and the products in the designed file system, and a big database for virtual reality (VR) data as an example.

Hardware Layer

Lamont HPC

CU HPC

Computing Services

XSEDE CLOUD COMPUTING SERVICES
READ MORE

GUI Portal – Data Download
DATA STORAGE
READ MORE

Data Services

Cloud Data Archiving