PGG | Data Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

DOI: *10.7916/D8VH5TJZ*
URL: *http://dx.doi.org/10.7916/D8VH5TJZ*

A Use Case of Big Data

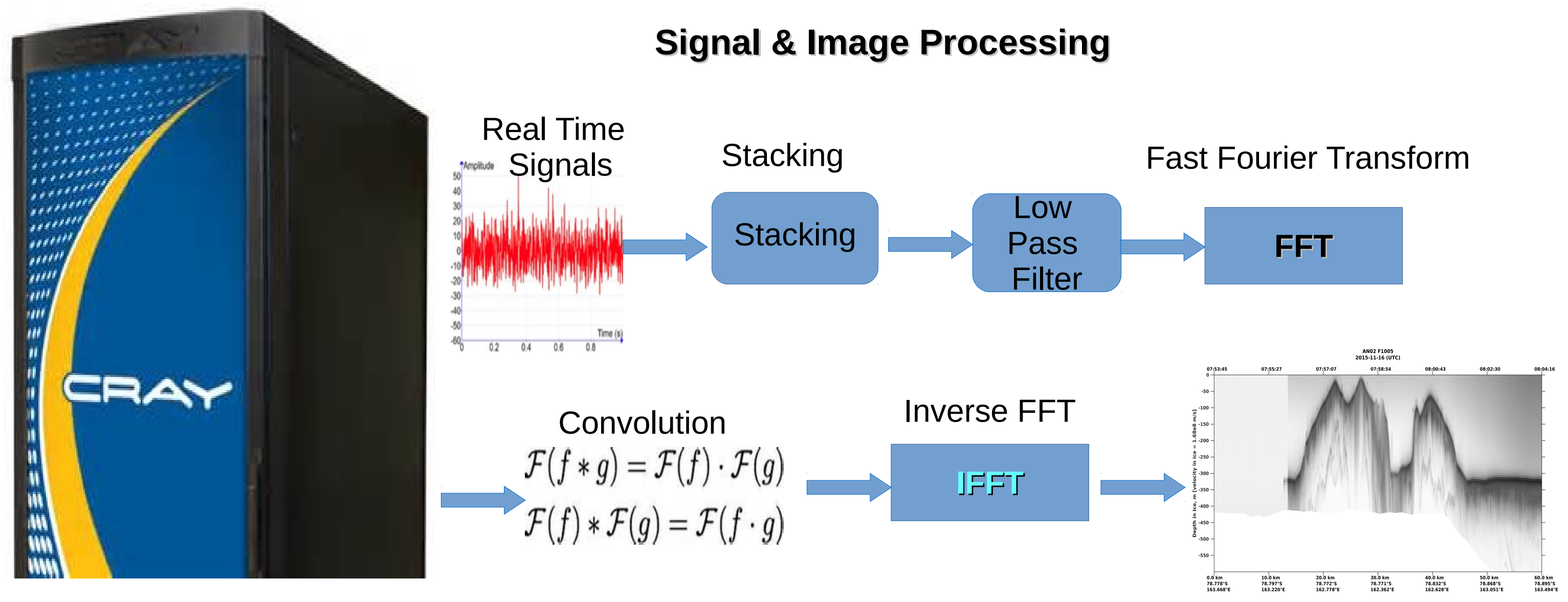
LingLing Dong ldong@ldeo.columbia.edu Nick Frearson nfre@ldeo.columbia.edu

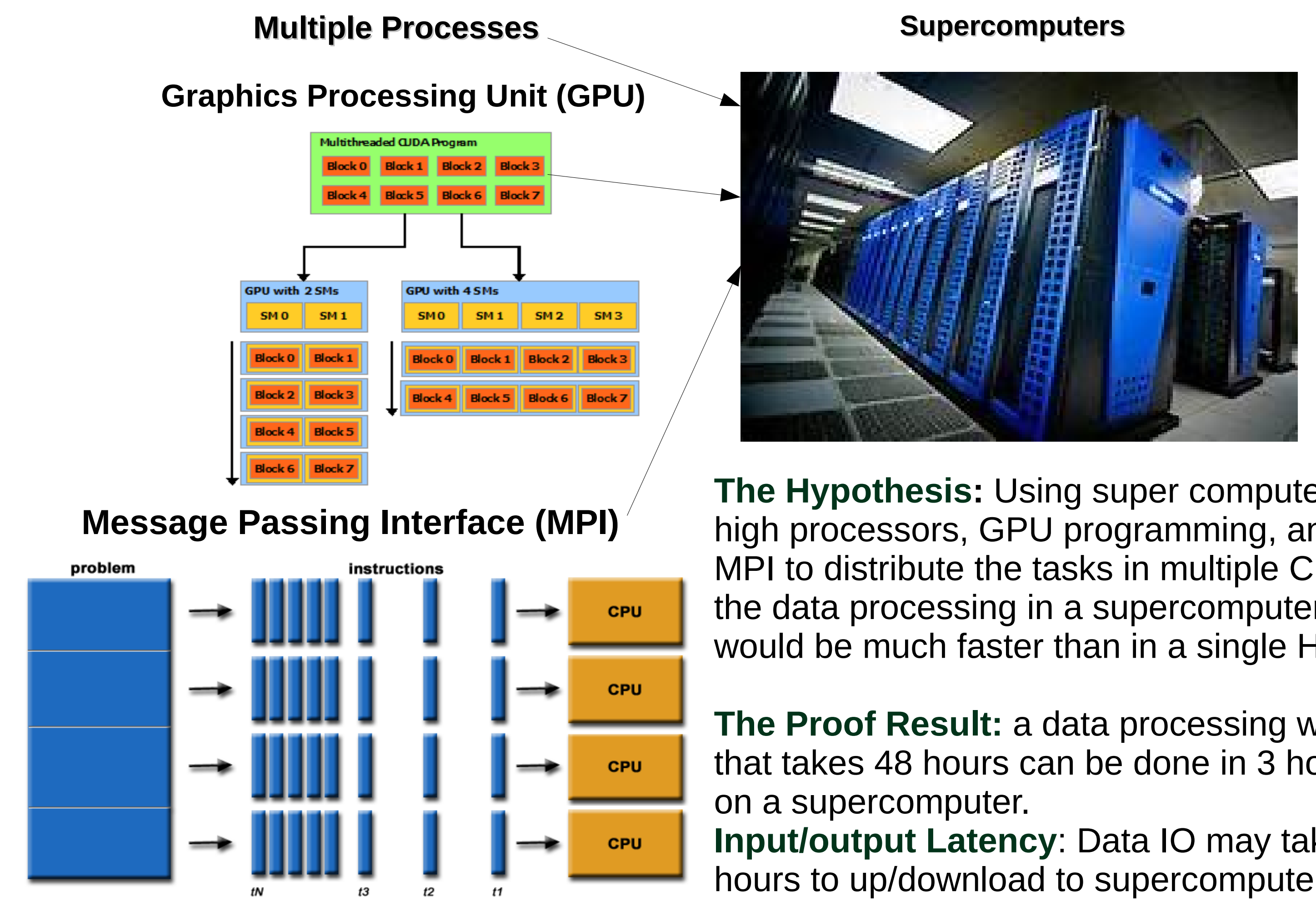
Case 1. Improve Performance with Supercomputer Services

Supercomputer Saves Data Processing Time

Reduced the operation time from 48 hours to 3 hours

Signal & Image Processing


CRAY
Real Time Signals
Stacking
Stacking
Low Pass Filter
Fast Fourier Transform
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



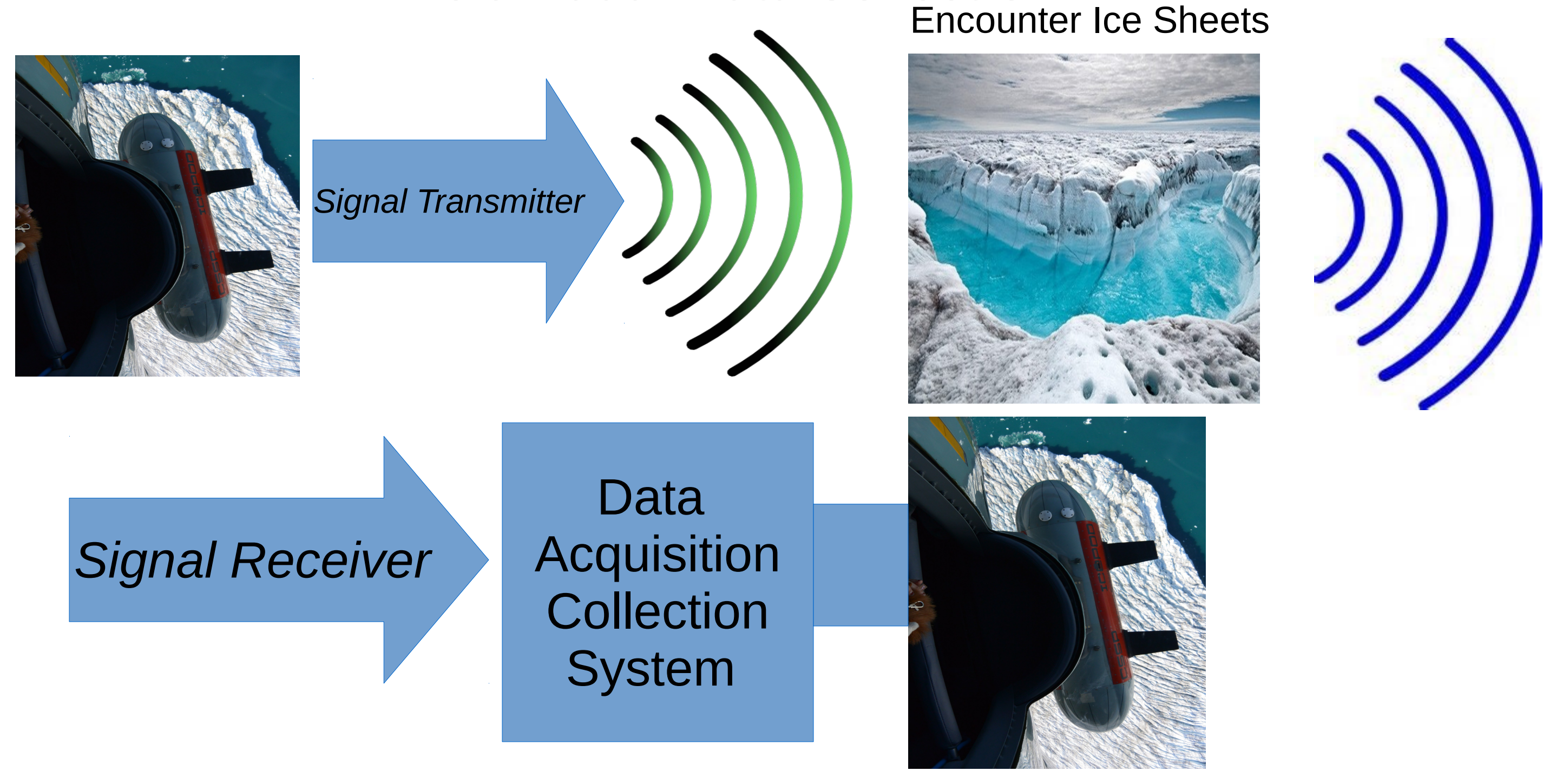
The Hypothesis: Using super computer's high processors, GPU programming, and MPI to distribute the tasks in multiple CPUs, the data processing in a supercomputer would be much faster than in a single HPC.

The Proof Result: a data processing work that takes 48 hours can be done in 3 hours on a supercomputer.

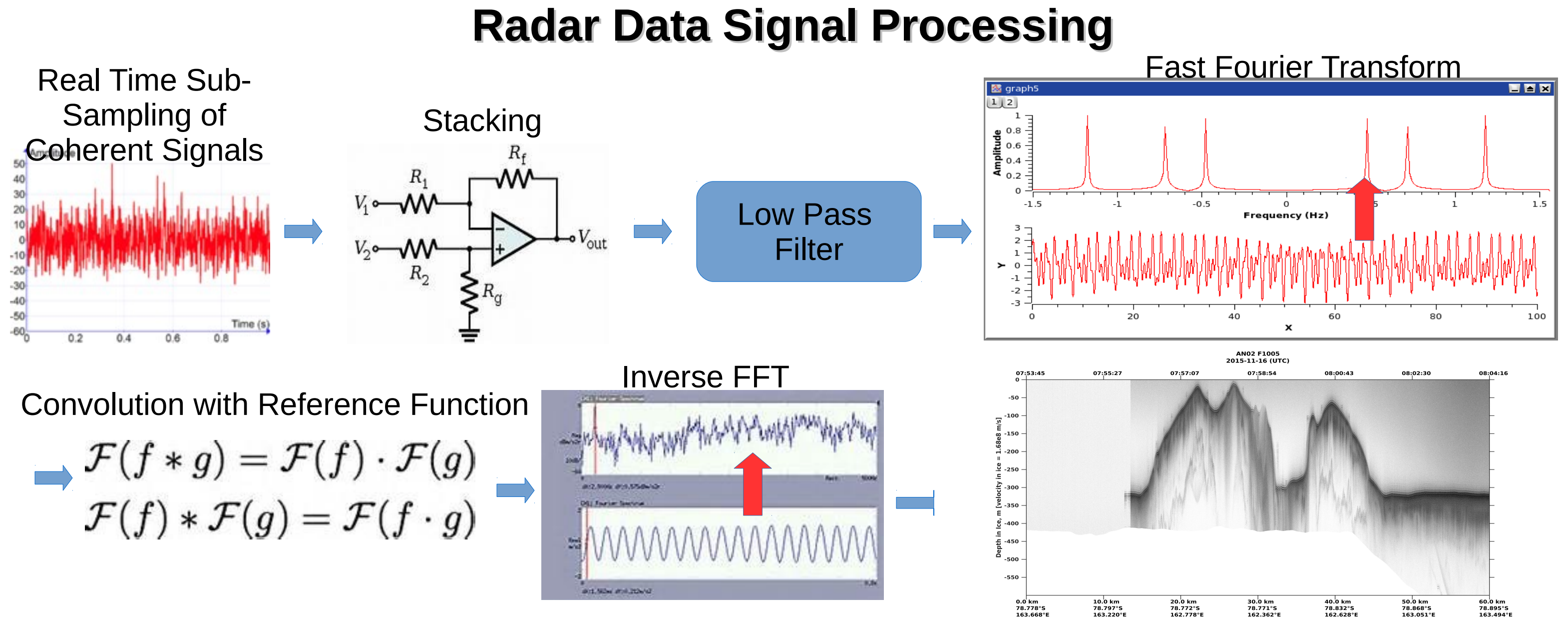
Input/output Latency: Data IO may take 2 hours to up/download to supercomputers.

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



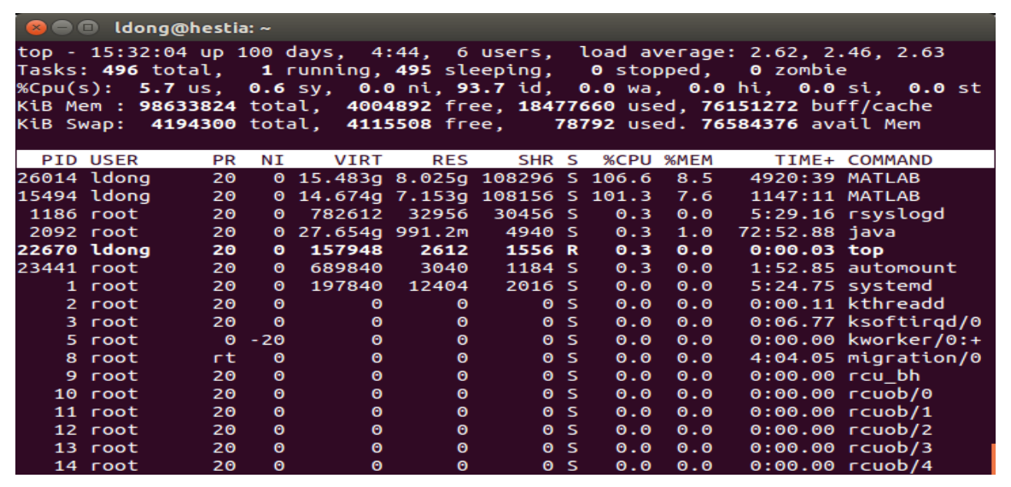
Radar Data Signal Processing


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
Frequency (Hz)

Case 2: Improve Performance in Big Data Operations

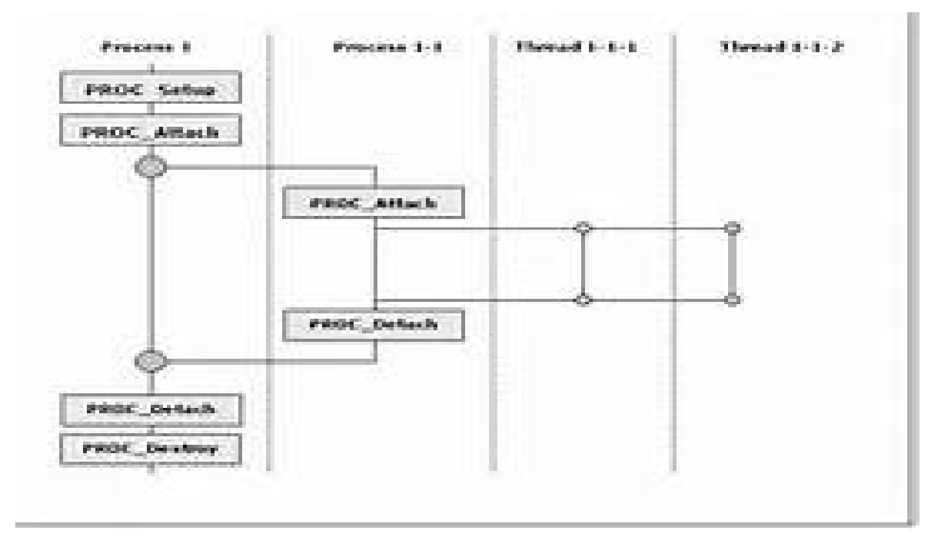
Resolved the Big Data Caused CPU Stress

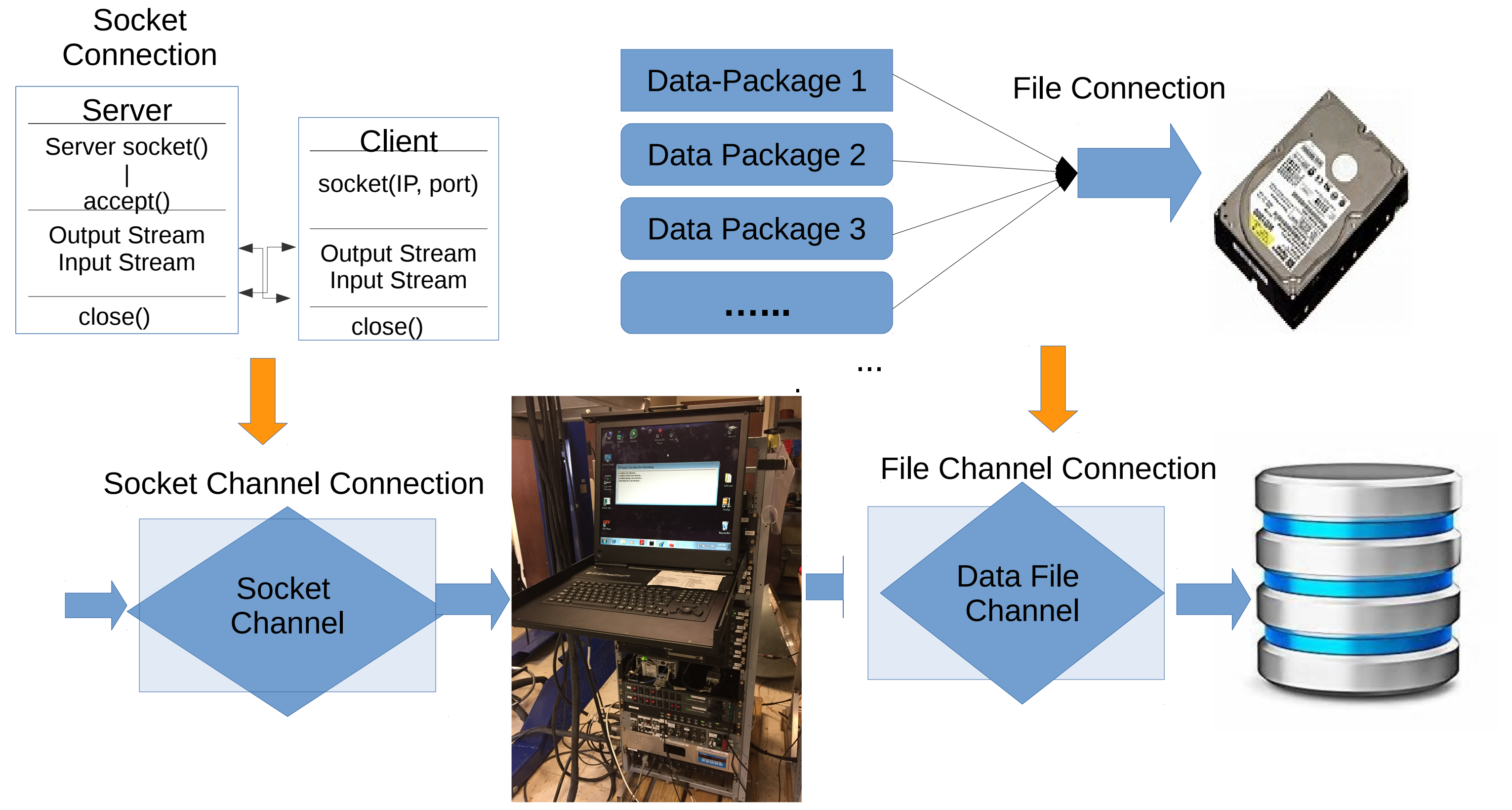
Reduced CPU usage from 100% to 15%



```
ldong@hestia: ~
top - 15:32:04 up 100 days,  4:44,  6 users,  load average: 2.62, 2.46, 2.63
Tasks: 496 total,   1 running, 495 sleeping,   0 stopped,   0 zombie
%Cpu(s):  5.7 us,  0.6 sy,  0.0 ni, 93.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 98633824 total,  4004892 free, 18477660 used, 76151272 buff/cache
KiB Swap:  4194300 total,  4115508 free,    78792 used. 76584376 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
26014 ldong     20   0 15.483g 8.025g 108296 S 106.6  8.5   4920:39 MATLAB
15494 ldong     20   0 14.674g 7.153g 108156 S 101.3  7.6   1147:11 MATLAB
 1186 root      20   0  782612  32956  30456 S   0.3  0.0   5:29.16 rsyslogd
 2092 root      20   0 27.654g 991.2m   4940 S   0.3  1.0  72:52.88 java
22670 ldong     20   0  157948   2612   1556 R   0.3  0.0   0:00.03 top
23441 root      20   0  689840   3040   1184 S   0.3  0.0   1:52.85 automount
    1 root      20   0  197840  12404   2016 S   0.0  0.0   5:24.75 systemd
    2 root      20   0       0      0      0 S   0.0  0.0   0:00.11 kthreadd
    3 root      20   0       0      0      0 S   0.0  0.0   0:06.77 ksoftirqd/0
    5 root       0 -20       0      0      0 S   0.0  0.0   0:00.00 kworker/0:+
    8 root      rt   0       0      0      0 S   0.0  0.0   4:04.05 migration/0
    9 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcu_bh
   10 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcuob/0
   11 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcuob/1
   12 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcuob/2
   13 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcuob/3
   14 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcuob/4
```





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

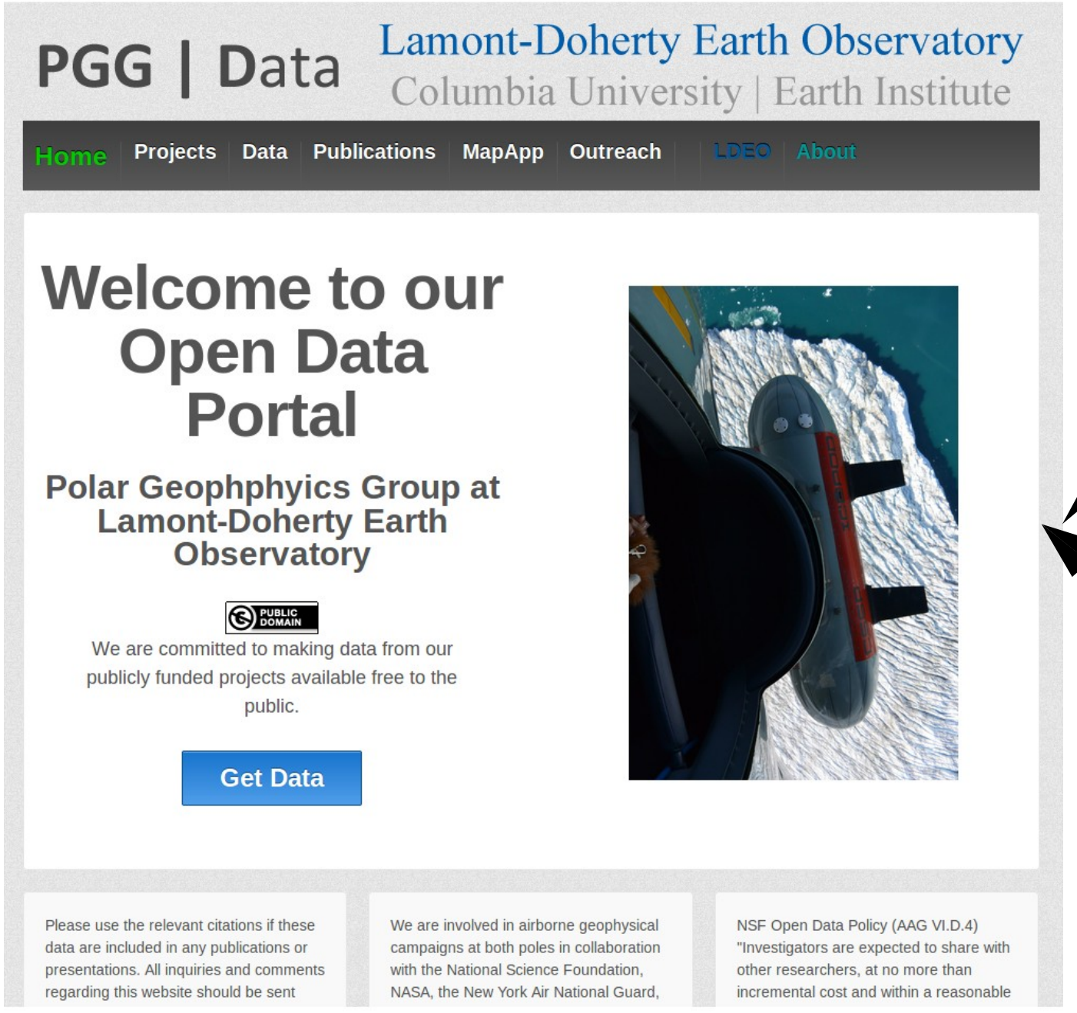
Data Apps | DAQ |Data Integration

Data Processing | Data Visualization| VR Data | Data Mining |Data Treasure||Data Acquisition|| SQL | NoSql

HARDWARE INFRASTRUCUTRE

Local Computers | Campus HPCs | XSEDE Cloud Services | Local Storage | XSEDE Storage ...

View Layer



PGG | Data Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

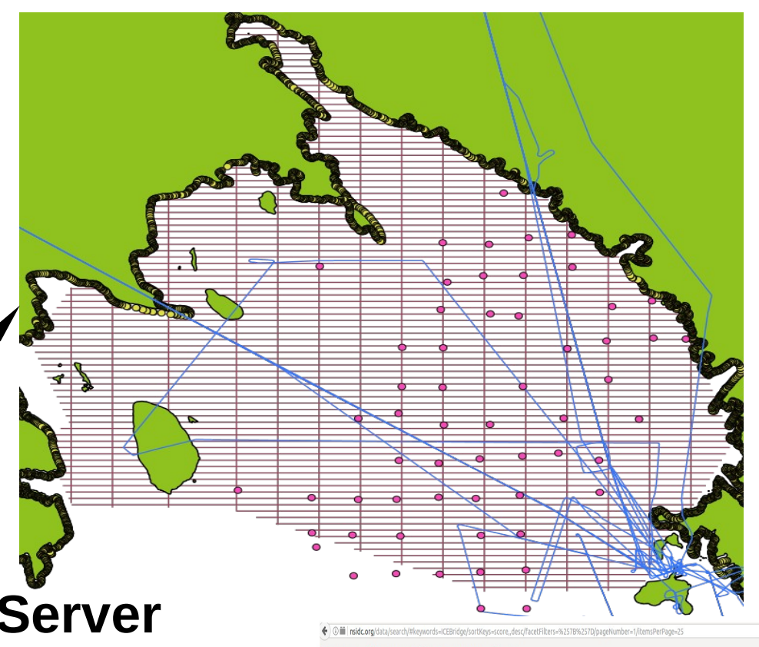
Home Projects Data Publications MapApp Outreach

Welcome to our Open Data Portal

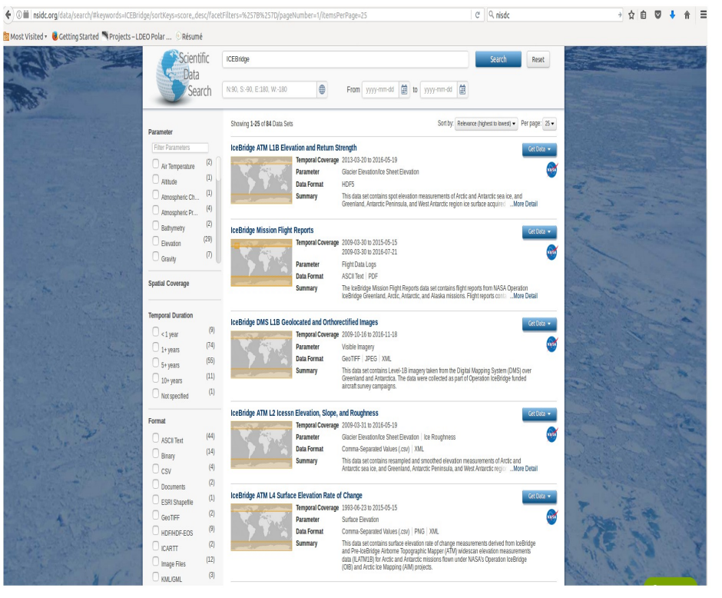
Polar Geophysics Group at Lamont-Doherty Earth Observatory

We are committed to making data from our publicly funded projects available free to the public.

Get Data

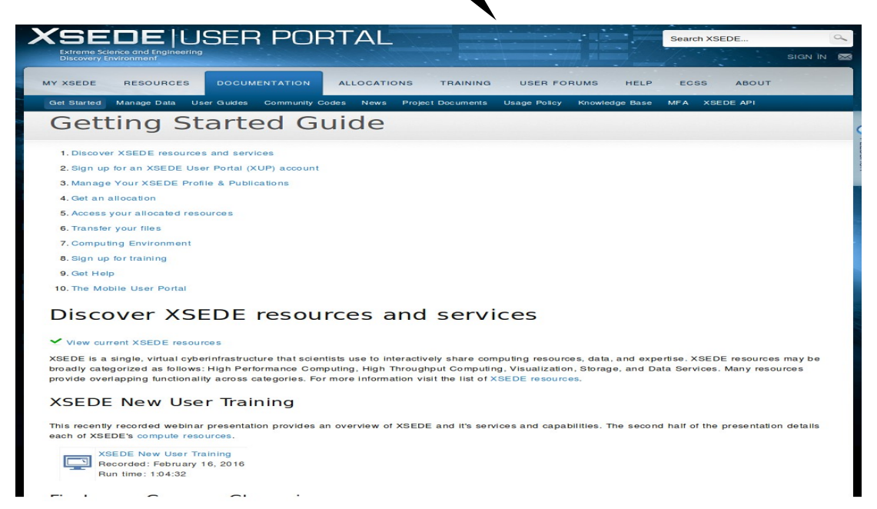


Map Server



Data Server

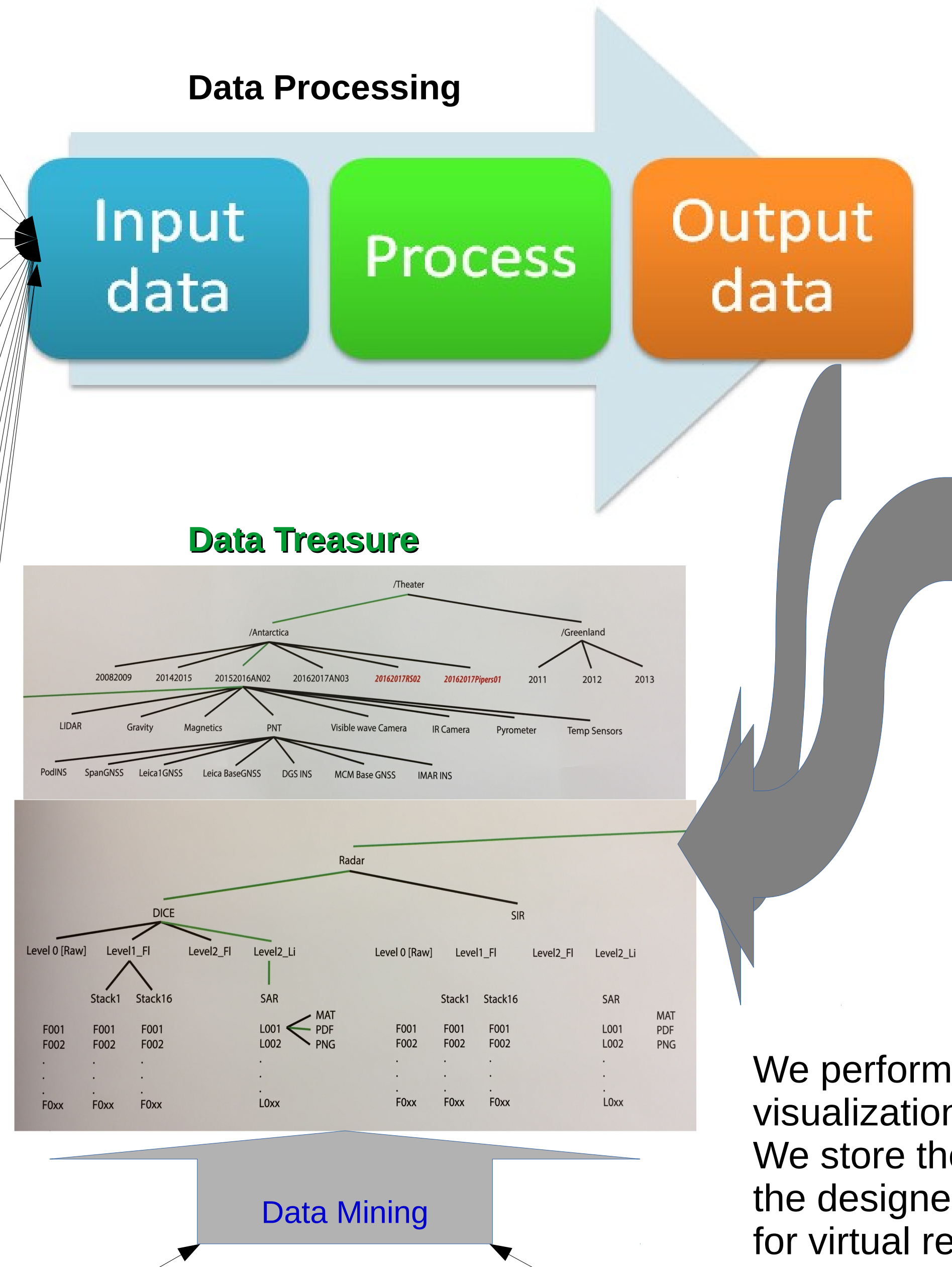
XSEDE Portal

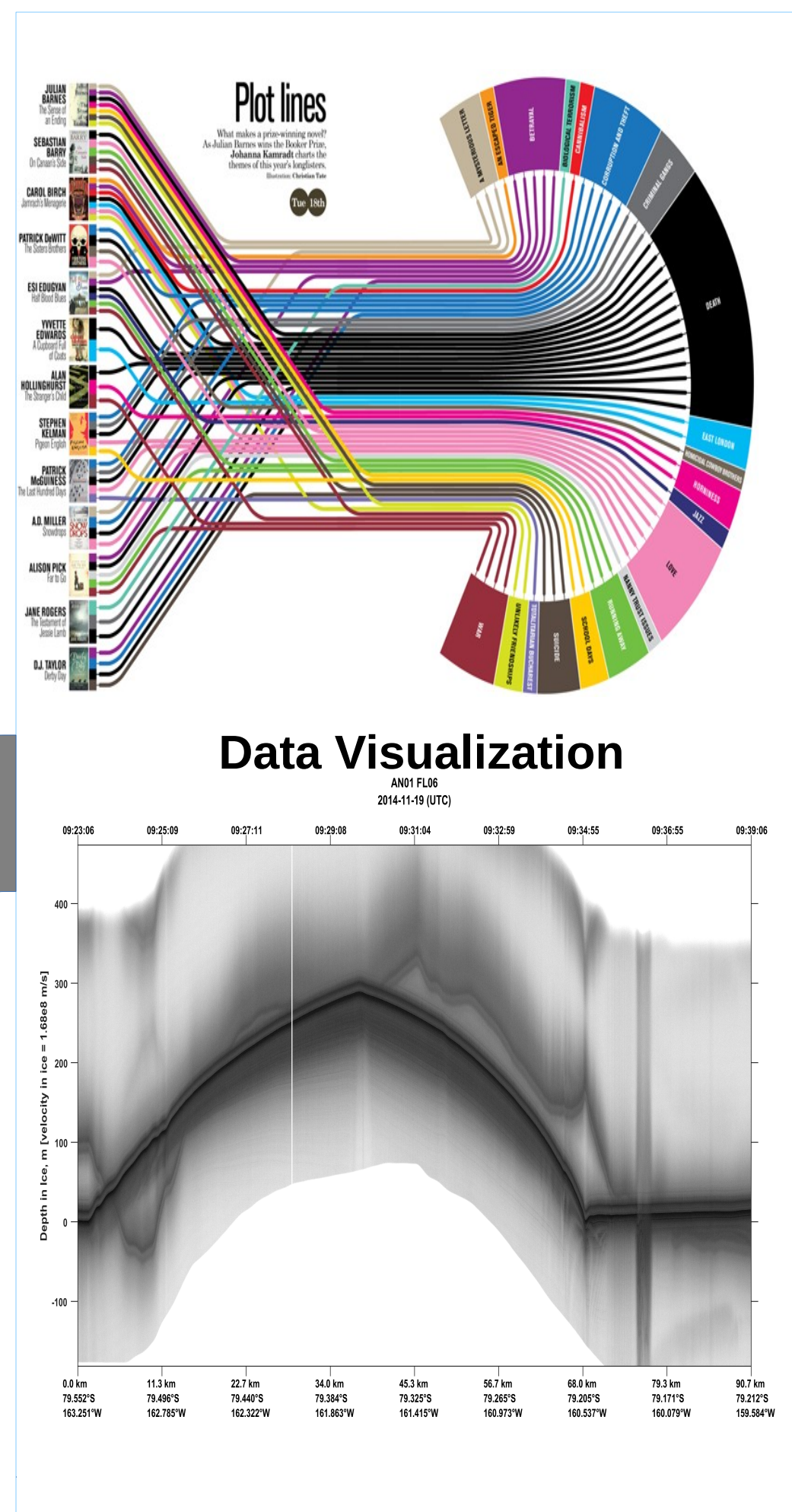


The Polar Geophysics Group web server, wonder.ldeo.columbia.edu, will host services from the map servers, the data publication servers, and our science portal in the XSEDE supercomputer center.

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 Layer

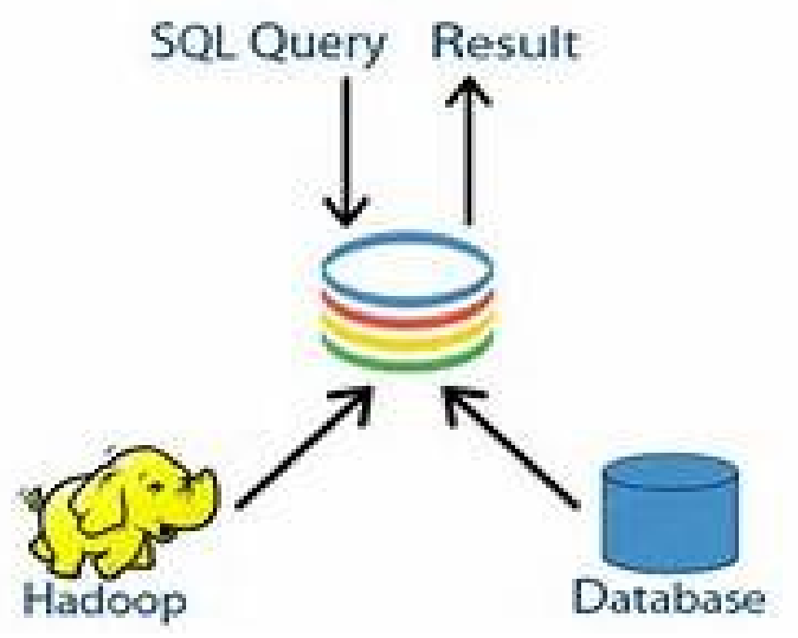




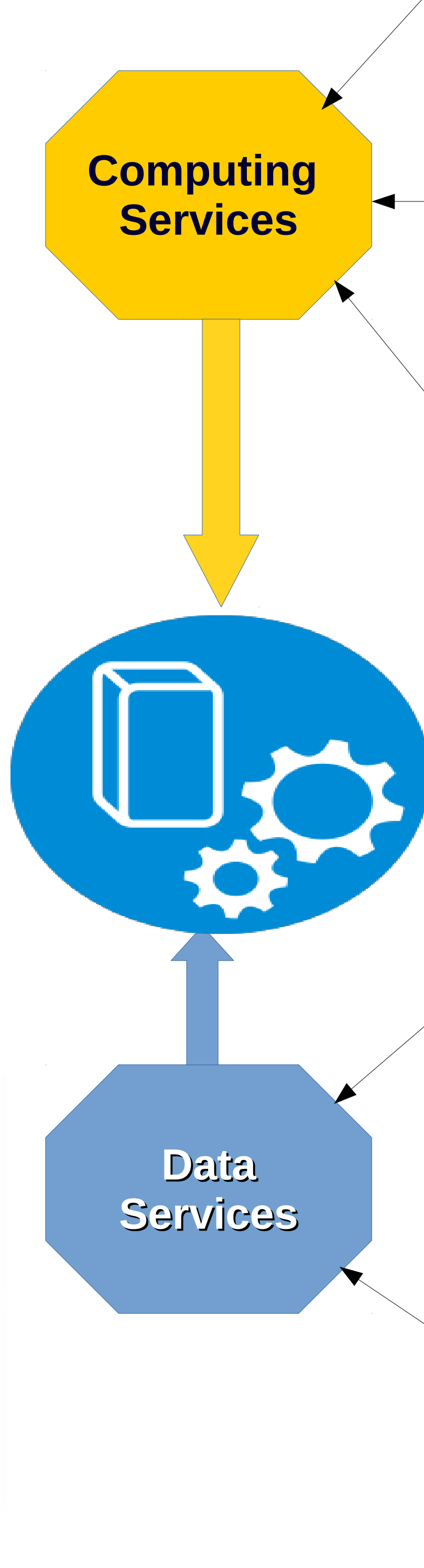
Data Layer

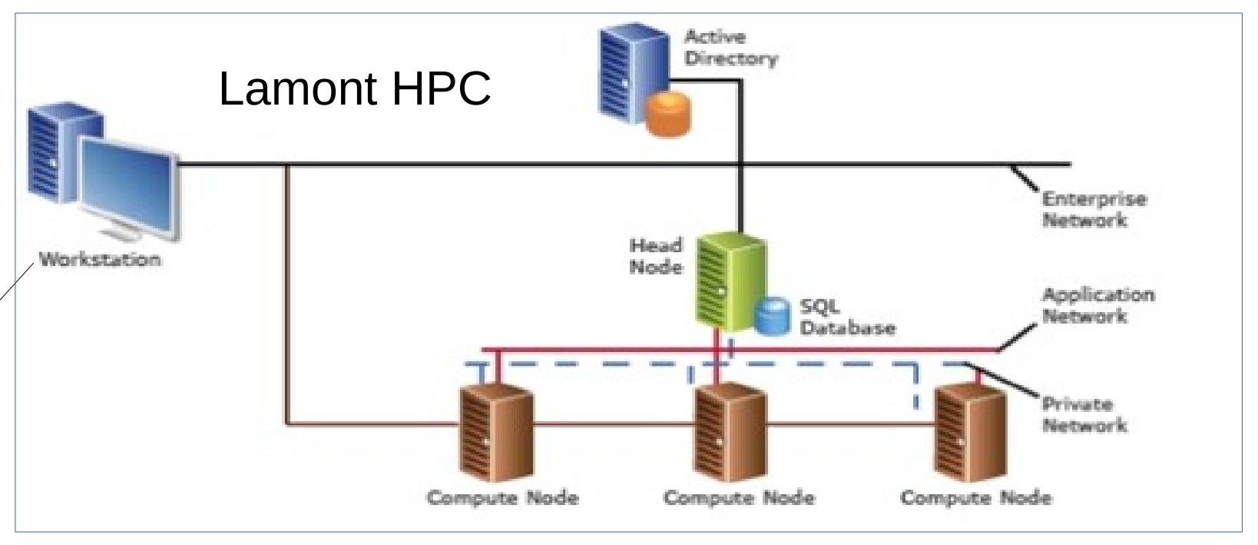
We perform data processing, data visualization, data mining etc.

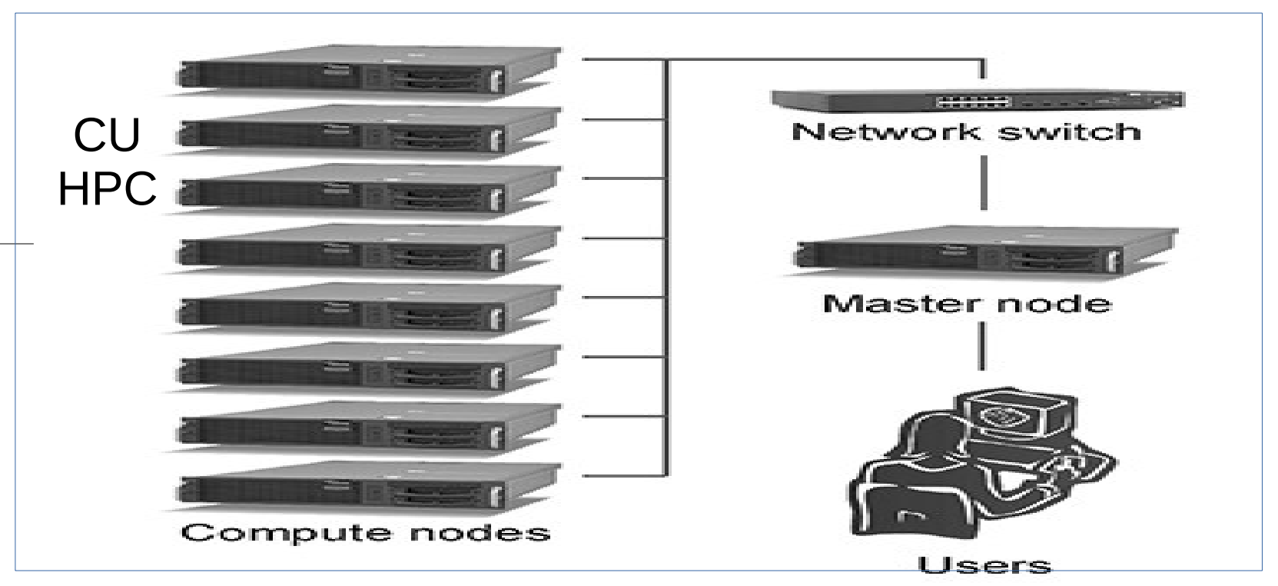
We store the meta data and the products in the designed file system, and a big database for virtual reality (VR) data as an example.



Hardware Layer







GUI Portal – Data Download