面临的问题/挑战

NVST-太阳观测数据高速存储

报 告 人：刘应波, 王锋
单　　位：中科院云南天文台
　　：云南省计算机技术应用重点实验室
日　　期：2013-11-16

NVST简要介绍

- **一米新真空红外太阳望远镜**[1]
  主要进行太阳高分辨率的太阳成像、光谱以及磁场观测，高精度地同时探测太阳光球、色球磁场及其动力学特征。

- **观测模式：**
  多通道、多波段

- **当前主力观测波段：**
  TiO-band(7058Å)
  G-band(4300Å)
  H-alpah(6562.8Å)

[1] http://fso.ynao.ac.cn/index.aspx

NVST数据基本情况

- **Camera**
  -Pco.2000, 4000 sCMOS ;
  -NEO Andor sCMOS ;

Channel	No. of Camera	Resolution (pixels)	Rate (fps)	Total Speed (MB/s)	Total Speed (GB/hr)
TiO-band(7058Å)	1	2560×2160	10	105	379
G-band(4300Å)	1	2560×2160	10	105	379
H-alpah(6562.8Å)	1	4008×2672	10	201	724
Total			**30**	**411**	**1482**

NVST未来数据挑战

- **增加观测波段**
  6503Å, 8452Å, 10803Å, …

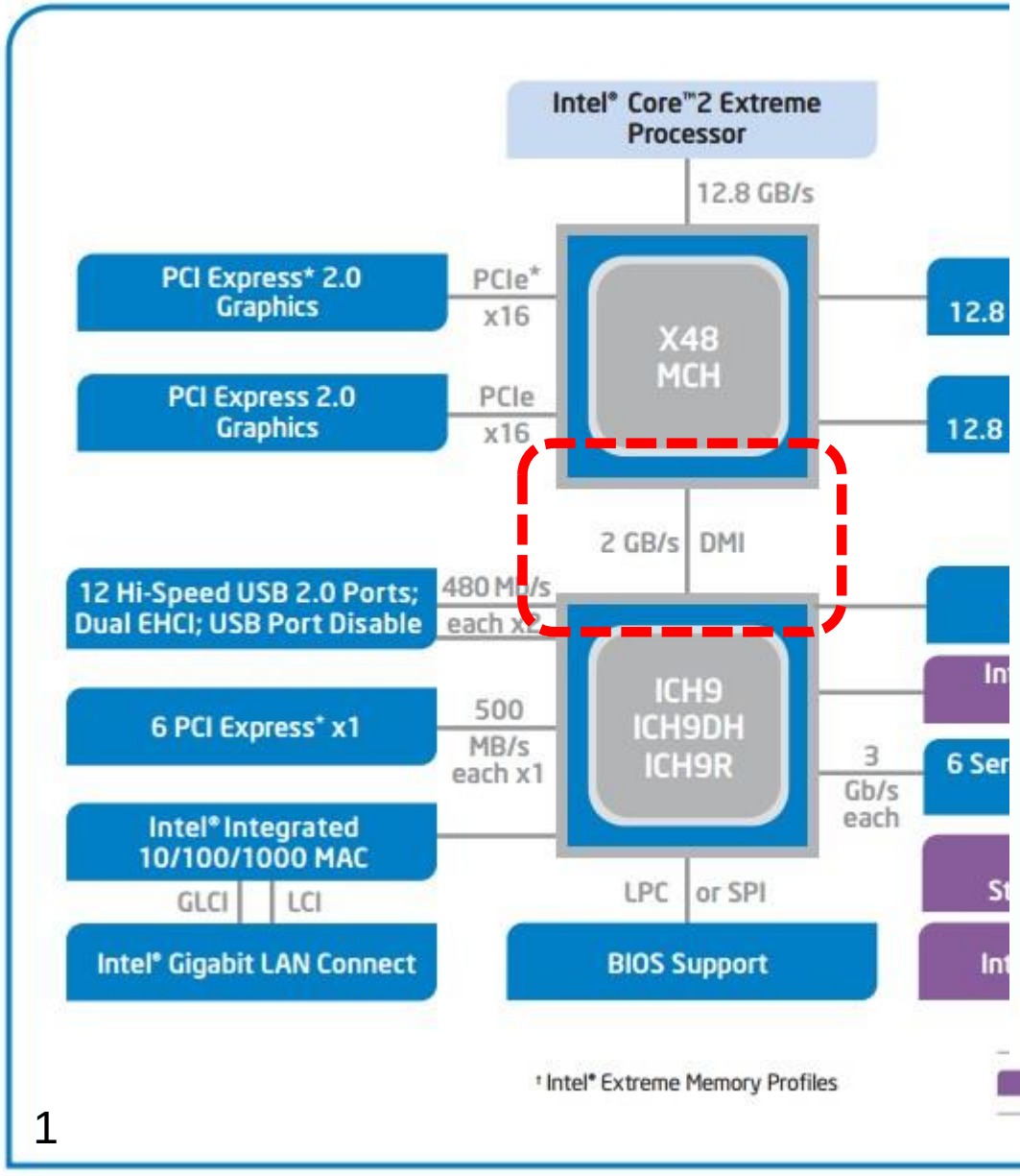
- **增加其他需求**
  增加即时数据计算需求, 例如，实时的数据计算等，并行的数据读写导致I/O需求翻倍。

- **提高采集帧速率**
  现在的5-10fps，期望20fps, 40fps, 60fps, 70fps, …
  例：在70fps的时候，存储落地带宽：
  H-alpah(6562.8Å)=1407MB/s,

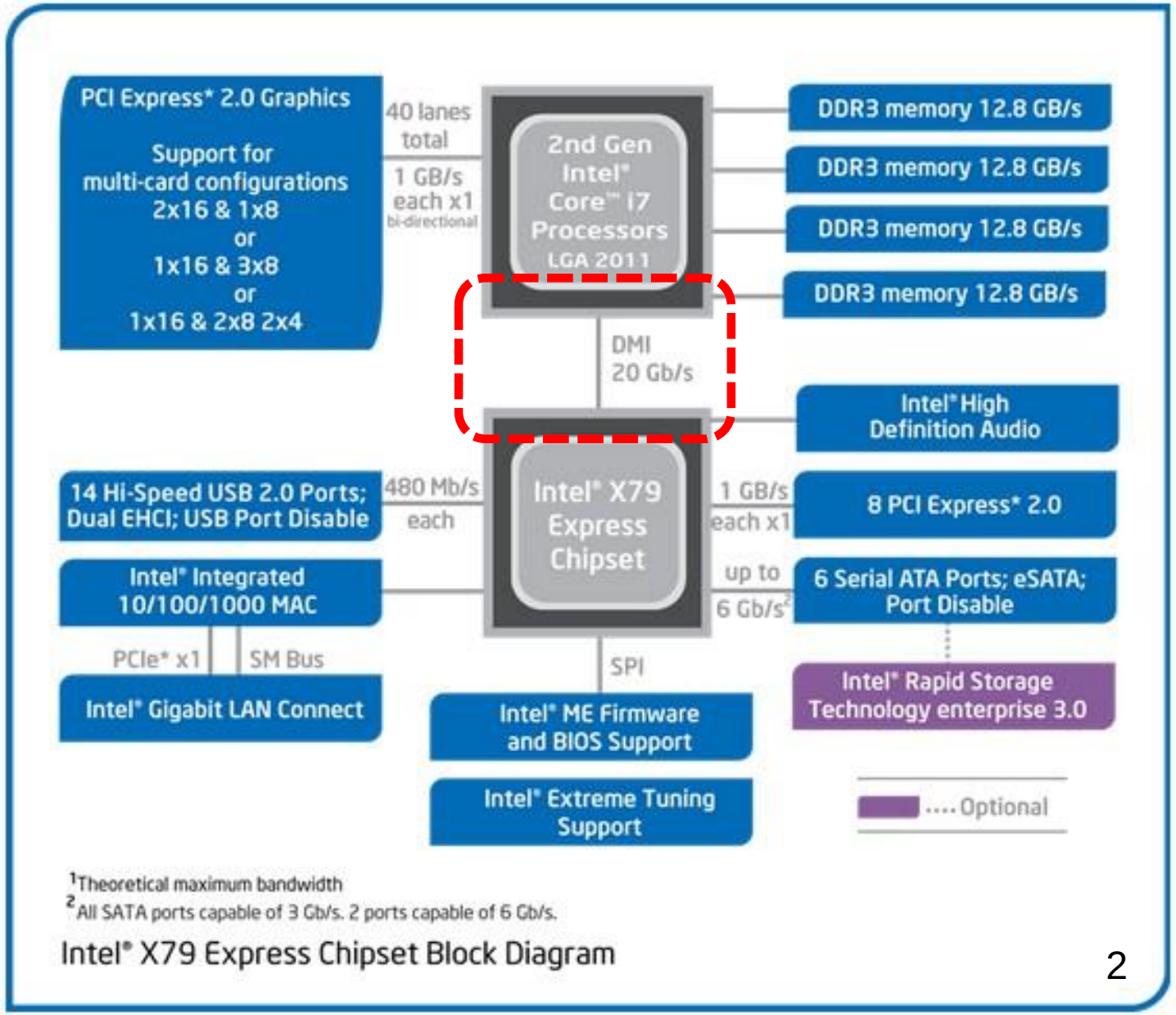
常用服务器存储相关部件带宽

- **硬盘**
  Barracuda XT 2TB（ST32000641AS）：持续数据传输率-138MB/秒，接口标准-SATA 6Gb/秒
- **网络**
  1GE-实际吞吐量小于125MB/s, 10GE-实际吞吐量小于1250MB/s
- **外围设备接口**
  PCI-E x1 双向400MB/s，x4 双向800MB/s，x8 双向1.6GB/s
- **南北桥总线**
  以Intel为例，Intel x48芯片组, DMI 单向1GB/s, X79芯片组, DMI 单向1.3GB/s

Intel® X48 Express Chipset Block Diagram



1



[1]Theoretical maximum bandwidth
[2]All SATA ports capable of 3 Gb/s. 2 ports capable of 6 Gb/s.

Intel® X79 Express Chipset Block Diagram

2

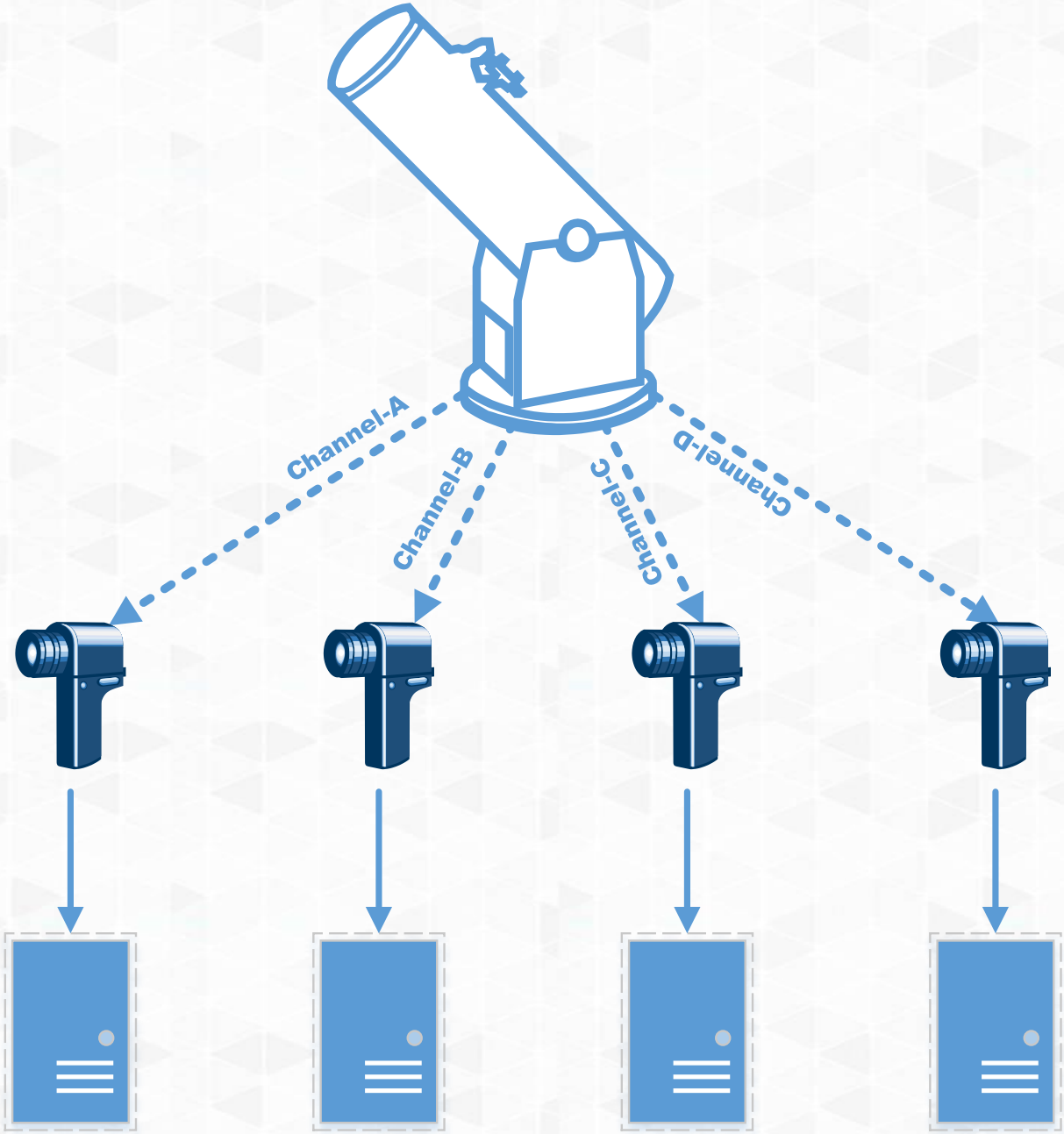
Intel® X79 Express Chipset Diagram

1. http://www.intel.com/Assets/PDF/prodbrief/319646.pdf
2.http://www.intel.com/content/www/xr/en/chipsets/performance-chipsets/x79-express-chipset-diagram.html

前置机数据采集模式

前置采集服务器

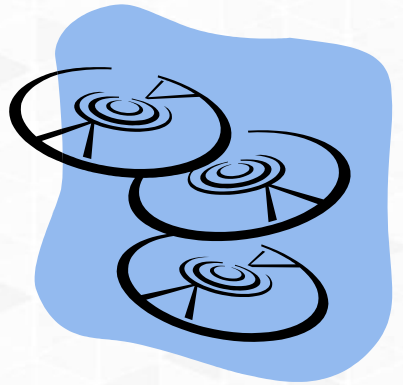
1. FITS 数据合成
2. 存储数据
3. 观测监控
...



NVST数据存储情况

以NVST为代表的太阳数据存储需求

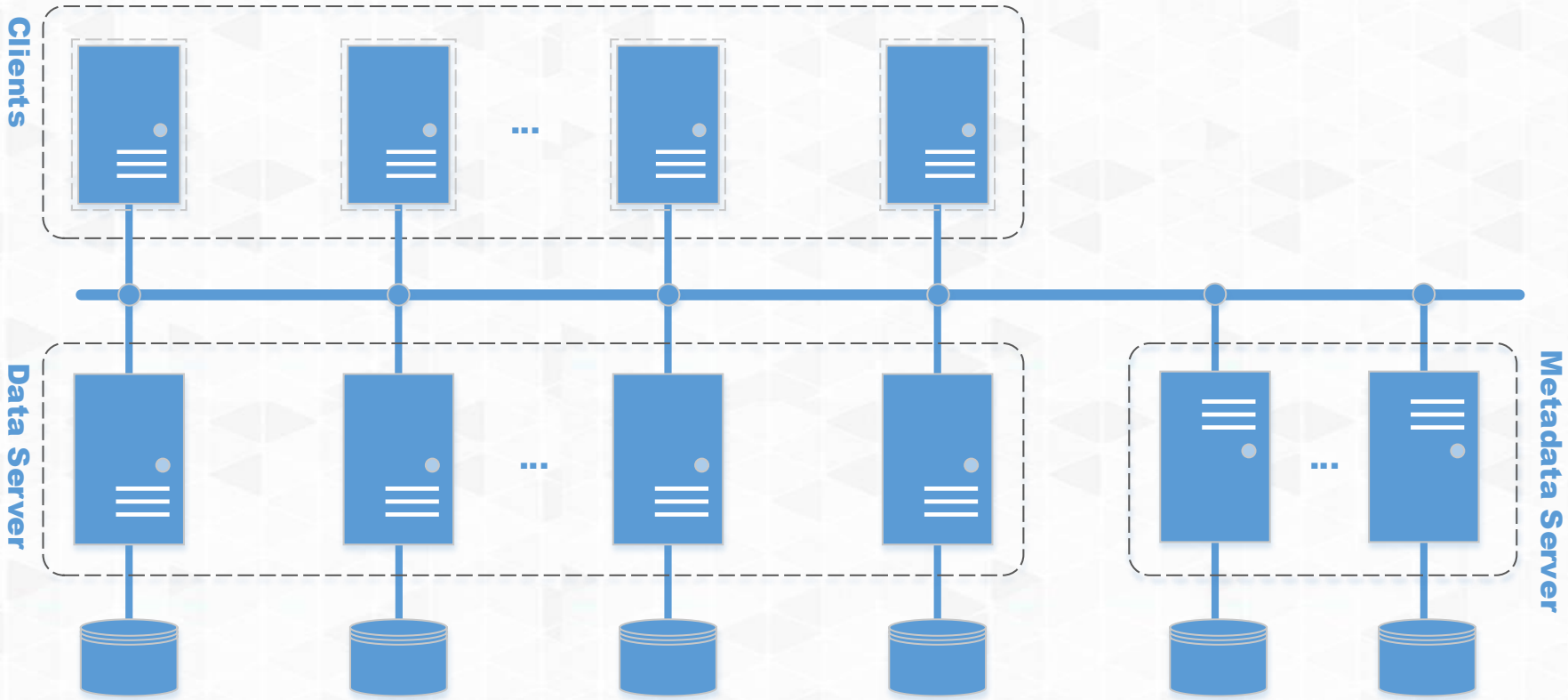
- **高性能数据存取**
  -支持高质量sCMOS观测数据
- **稳定的长时间数据读写**
  -支持全天候连续数据观测
- **顺序I/O和随机I/O**
  -流式观测数据存储，即时数据查看，实时计算
- **高扩展性**
  -满足多通道，多波段，海量数据存储
- **可管理、可维护**

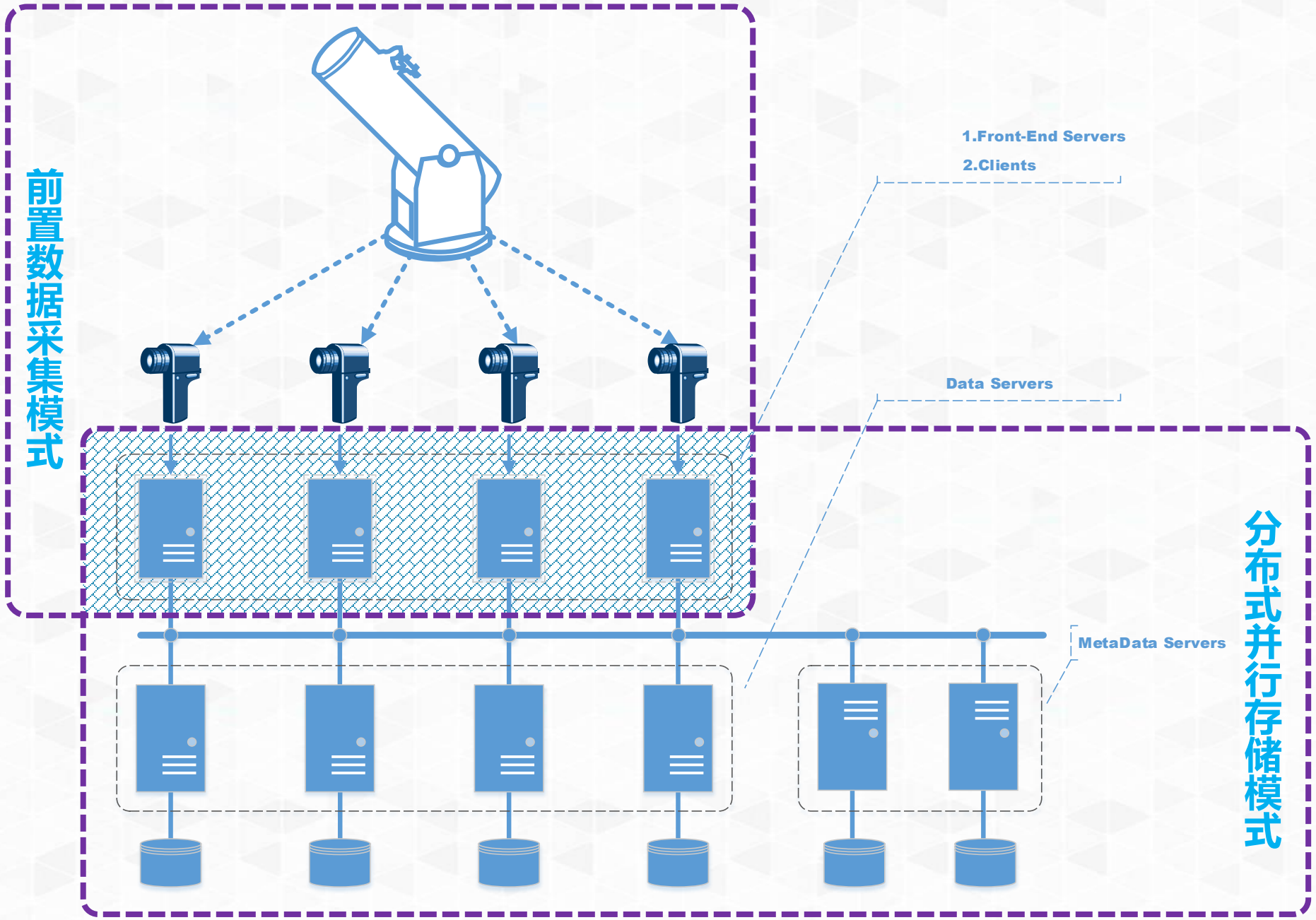


提高存储性能两种思想

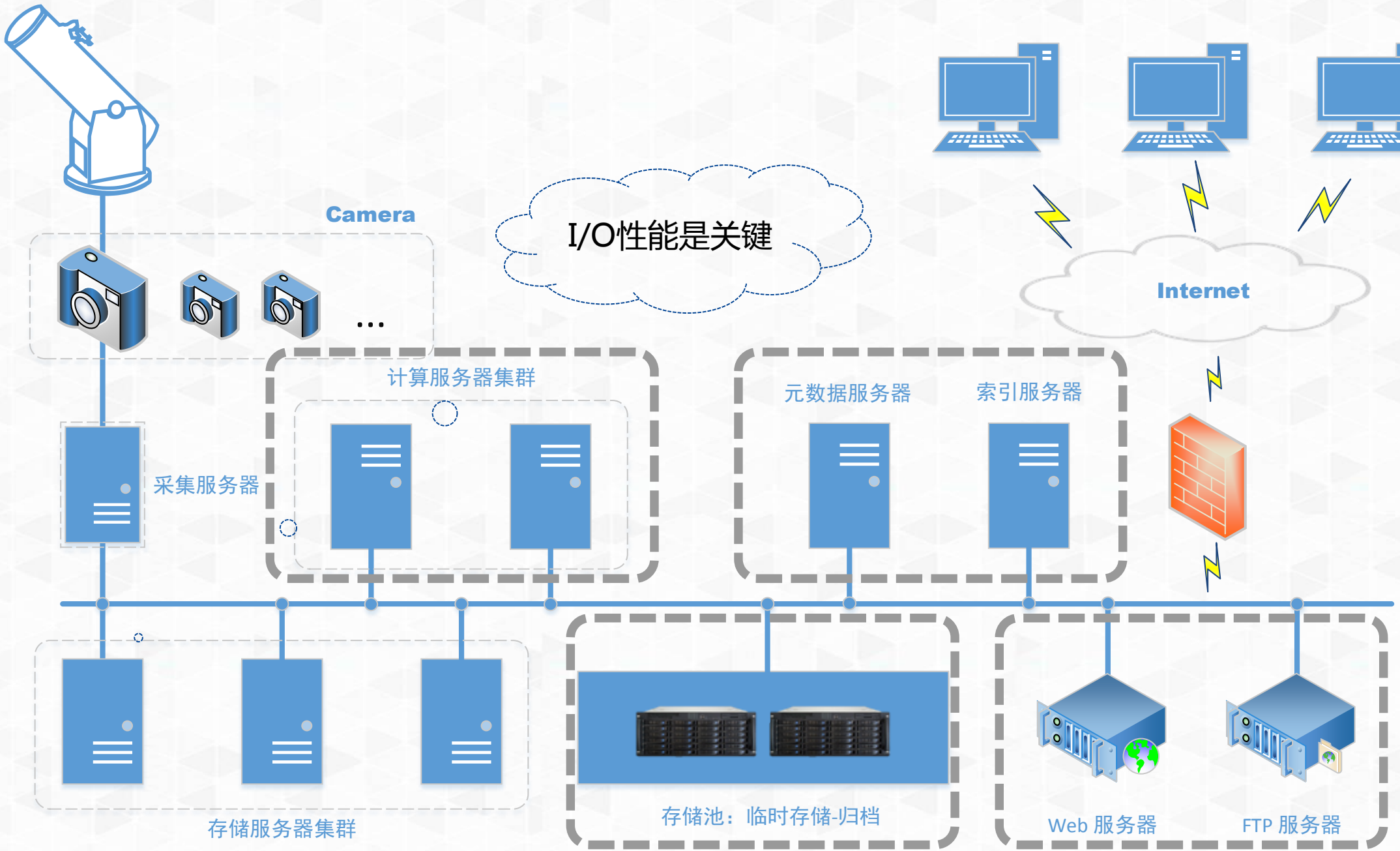
- **纵向扩展(Scale-up)**
  升级硬件、软件资源，例如：
  -机械硬盘换固态硬盘
  -内存容量，性能升级
  -1GE->10GE/Infiniband
  -…
- **横向扩展(Scale-out)**
  克服了物理机架和模块的限制，例如：
  -分布式并行存储（商业和开源产品）

分布式并行存储

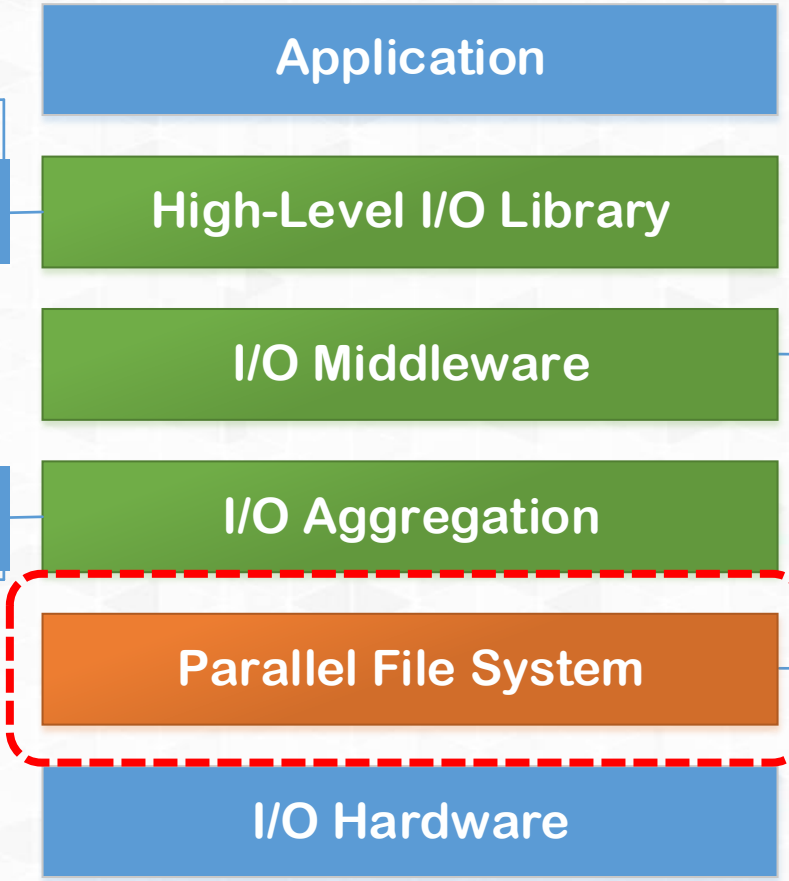



前置数据采集模式
1.Front-End Servers
2.Clients
Data Servers
MetaData Servers
分布式并行存储模式

NVST 存储结构图


Camera
...
I/O性能是关键
Internet
计算服务器集群
采集服务器
元数据服务器
索引服务器
存储服务器集群
存储池：临时存储-归档
Web 服务器
FTP 服务器

The I/O 软件栈[1]



High-Level I/O Library
Maps application abstractions onto storage abstractions and provides data portability

HDF5，Parallel netCDF，ADIOS

I/O Middleware
Organizes accesses from many processes, especially those using collective I/O

MPI-IO

I/O Aggregation
Bridges between app. Tasks and storage system and combines operations to reduce "randomness"of I/O

IBM ciod

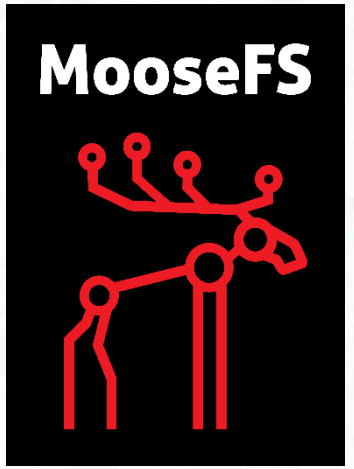
Parallel File System
Maintains logical space and provides efficient access to data

PVFS, PanFS,GPFS,Lustre

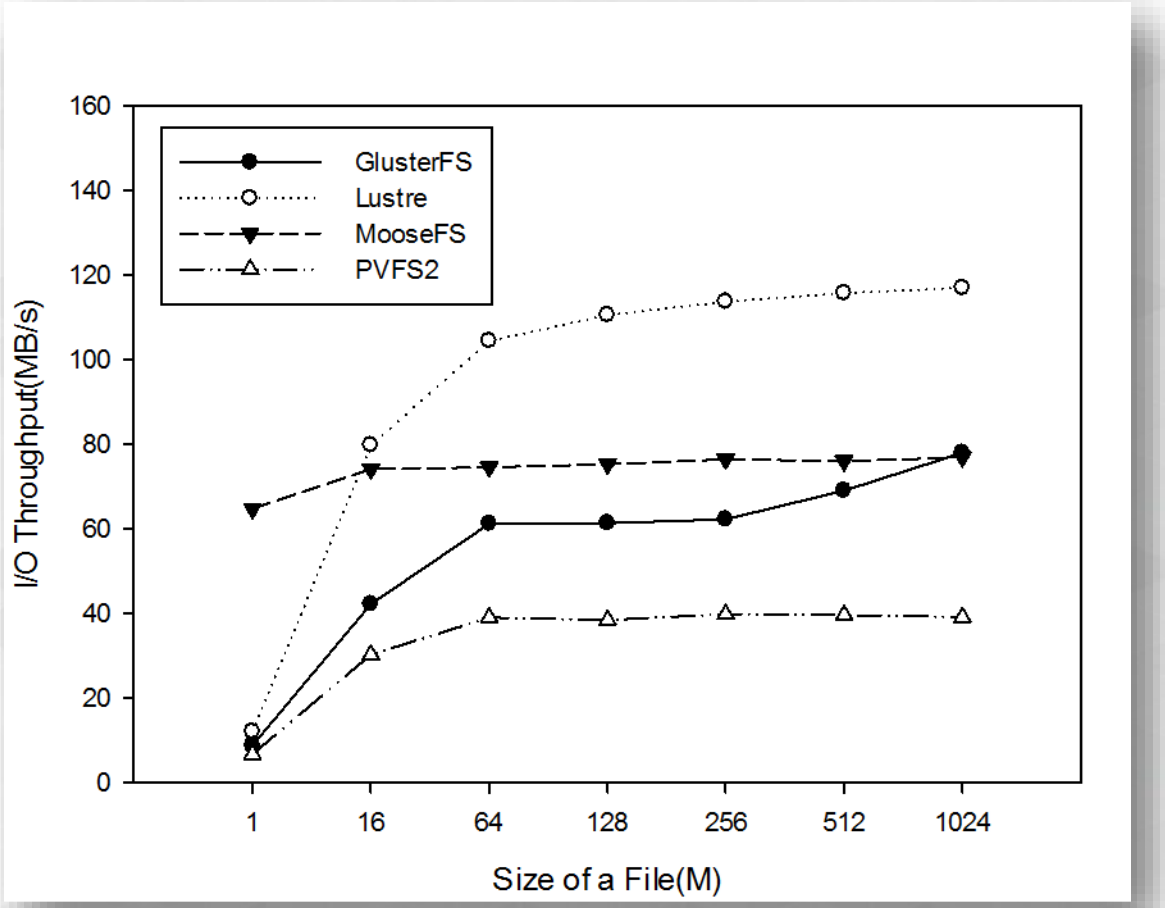
[1]R. Latham, C. Daley, W. Liao, K. Gao, R. Ross, A. Dubey, and A. Choudhary, "A case study for scientific I/O: improving the FLASH astrophysics code," *Computational Science & Discovery,* vol. 5, no. 1, pp. 015001, 2012.

分布式、并行文件系统

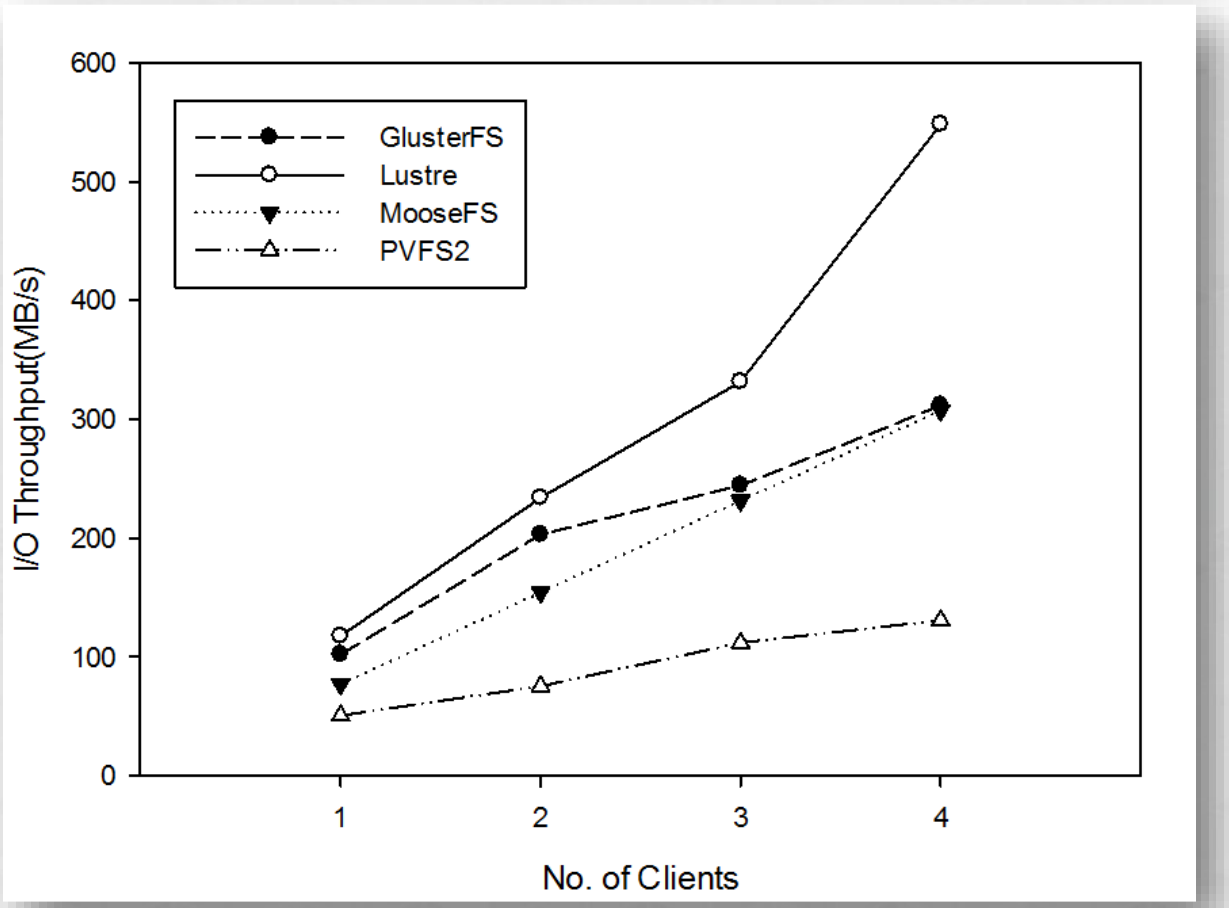
- 文件系统
  透明使用
- 数据存储容量和I/O性能横向扩展
- 维护和管理麻烦，需要专门支持团队



分布式并行存储的性能和扩展性



单通道性能



64M 多通道性能扩展性

Lustre典型应用环境

典型应用环境

1. Capture Environment
2. Parallel Environment

1.	Single stream with large data blocks operating in half duplex mode	
2.	Single stream with large data blocks operating in full duplex mode	
3.	Multiple streams with large data blocks operating in full duplex mode	
4.	Extreme file creation rates	Capture Environment
5.	Checkpoint/restart with large I/O requests	Parallel Environment
6.	Checkpoint/restart with small I/O requests	
7.	Checkpoint/restart large file count per directory large I/Os	
8.	Checkpoint/restart large file count per directory small I/Os	
9.	Walking through directory trees	
10.	Parallel walking through directory trees	
11.	Random stat() system call to files in the file system (one process)	
12.	Random stat() system call to files in the file system (multiple proc's)	
13.	Small block random I/O to multiple files	
14.	Small block random I/O to a single file	

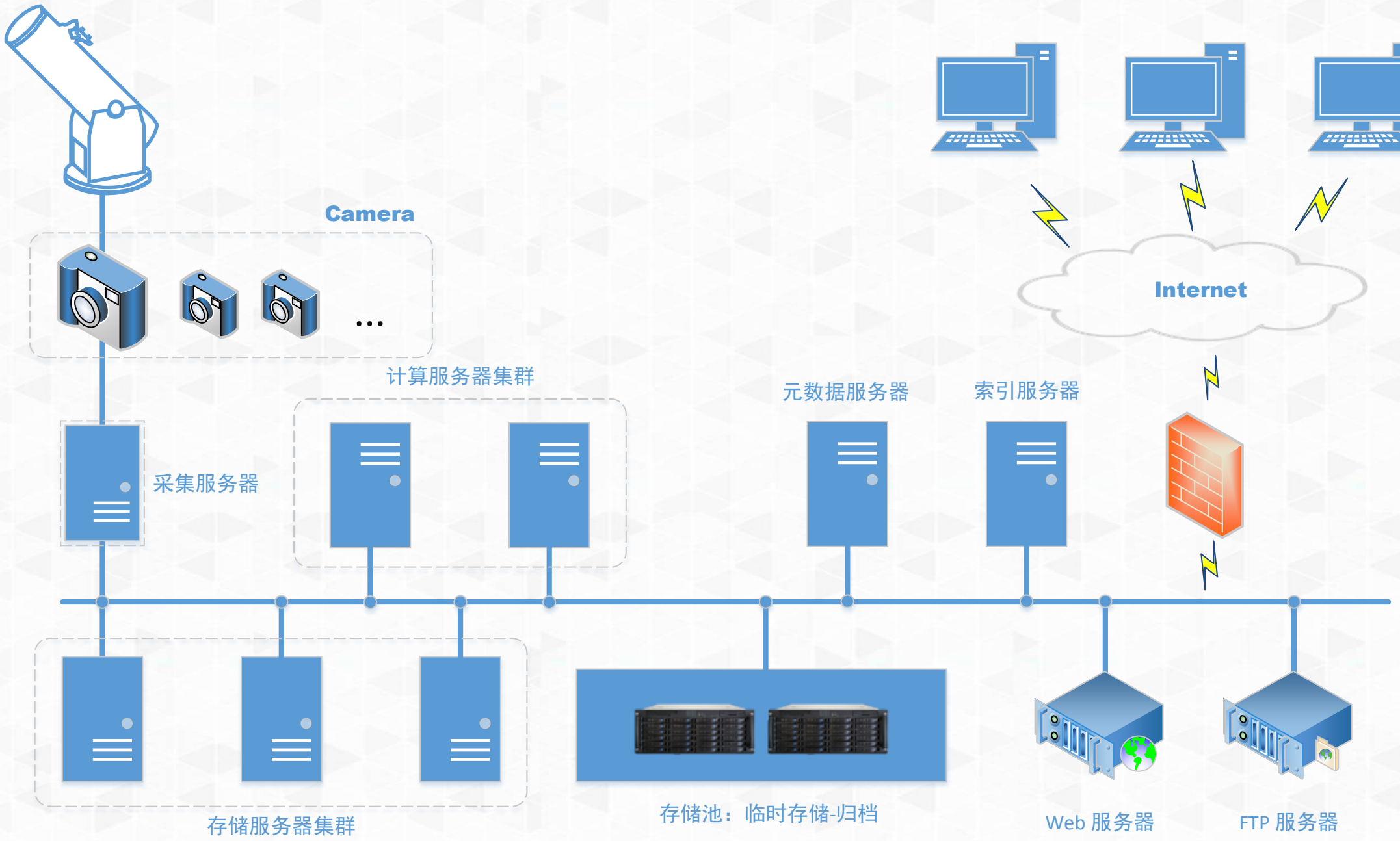
1

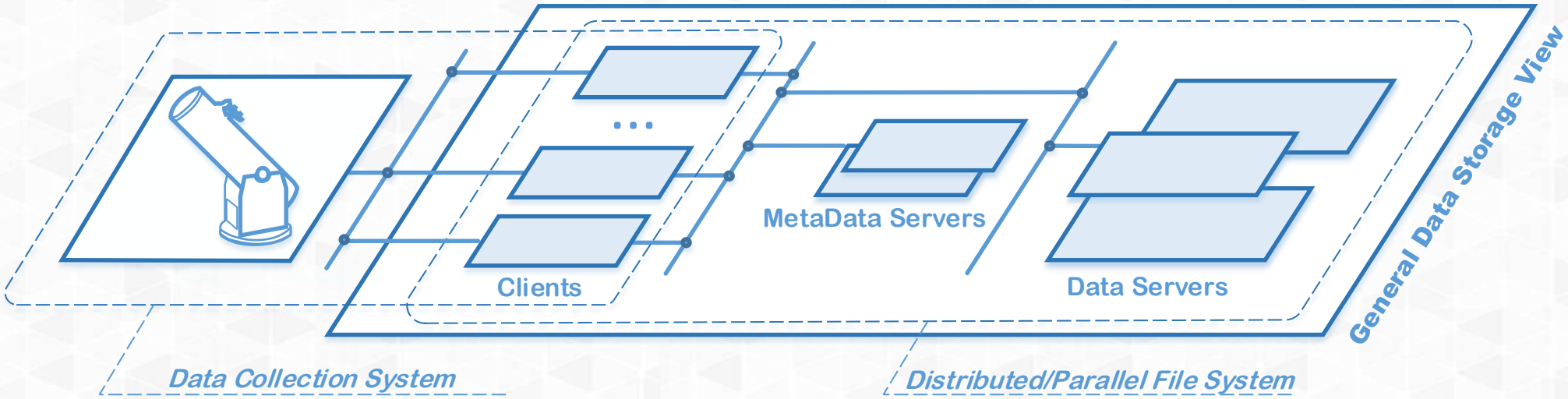
Chroma™

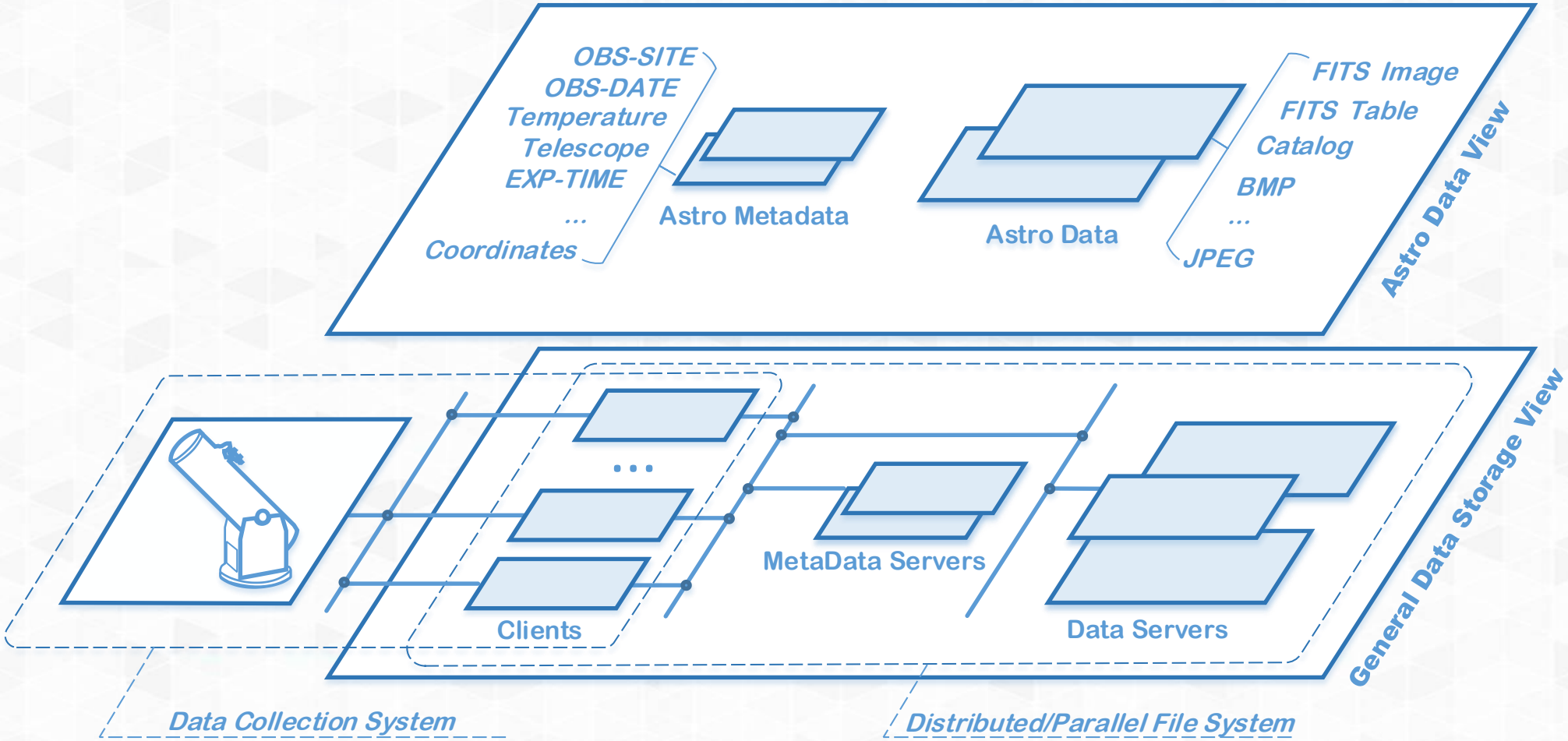
没有开源

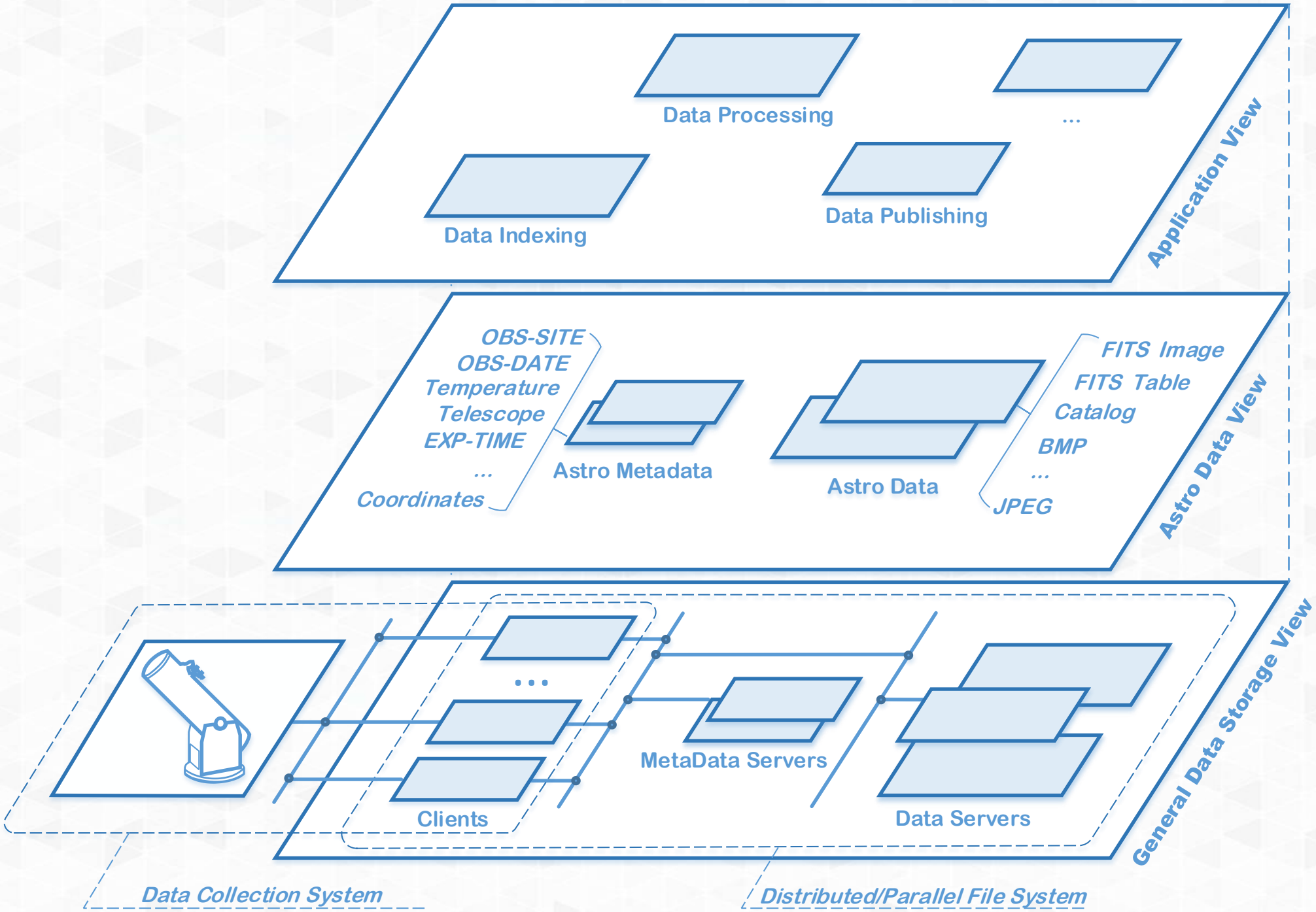
[1] http://www.opensfs.org/wp-content/uploads/2011/11/HPCS-IO-Scenarios-for-OBL-01.pdf

NVST 存储结构图


Camera
...
Internet
计算服务器集群
采集服务器
元数据服务器
索引服务器
存储服务器集群
存储池：临时存储-归档
Web 服务器
FTP 服务器


Clients
MetaData Servers
Data Servers
General Data Storage View
Data Collection System
Distributed/Parallel File System


OBS-SITE
OBS-DATE
Temperature
Telescope
EXP-TIME
...
Coordinates
Astro Metadata
Astro Data
FITS Image
FITS Table
Catalog
BMP
...
JPEG
Astro Data View
Clients
MetaData Servers
Data Servers
General Data Storage View
Data Collection System
Distributed/Parallel File System


Data Processing
...
Data Indexing
Data Publishing
Application View
OBS-SITE
OBS-DATE
Temperature
Telescope
EXP-TIME
...
Coordinates
Astro Metadata
Astro Data
FITS Image
FITS Table
Catalog
BMP
...
JPEG
Astro Data View
MetaData Servers
Clients
Data Servers
General Data Storage View
Data Collection System
Distributed/Parallel File System

总结

- **分布式并行是解决当前NVST高速大数据存储的有效方式;**
- **并行读写、读写分离、分级存储;**
- 现在的分布式并行的存储系统属于通用系统，可能需要一些定制的功能;

定位：太阳观测领域存储

元数据：带外方式

主要关注

性能

扩展

安全

容量扩展

IO扩展

Scale Out

流式数据存储

数据分片安全

冗余策略

数据校验策略

数据很大的情况下，冗余占用双倍甚至更多存储空间，类似校验码校验数据完整性

1. 应用考虑

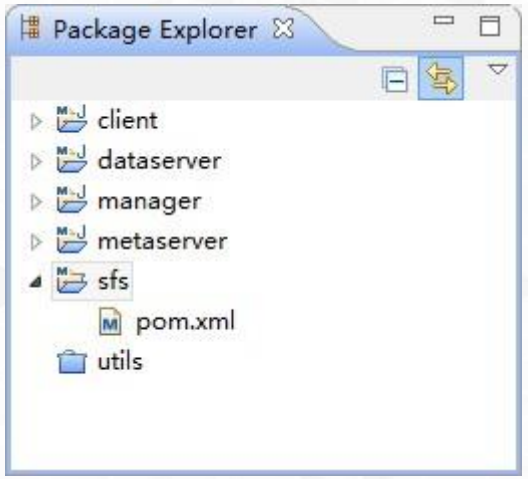
-直接使用现有开源系统
-现有系统上功能增减
-重新实现存储系统

2. 4月份分布式存储原型

libpaxos
libevent
leveldb

https://github.com/brianzf/GlobalAstroTable

Solar File Storage(SFS)



谢谢