sodacon 2020

DATA CONNECTED

CURVE：A NEXT-GEN HIGH-PERFORMANCE DISTRIBUTED STORAGE SYSTEM

Xiaocui Lee，NetEase Digitasail

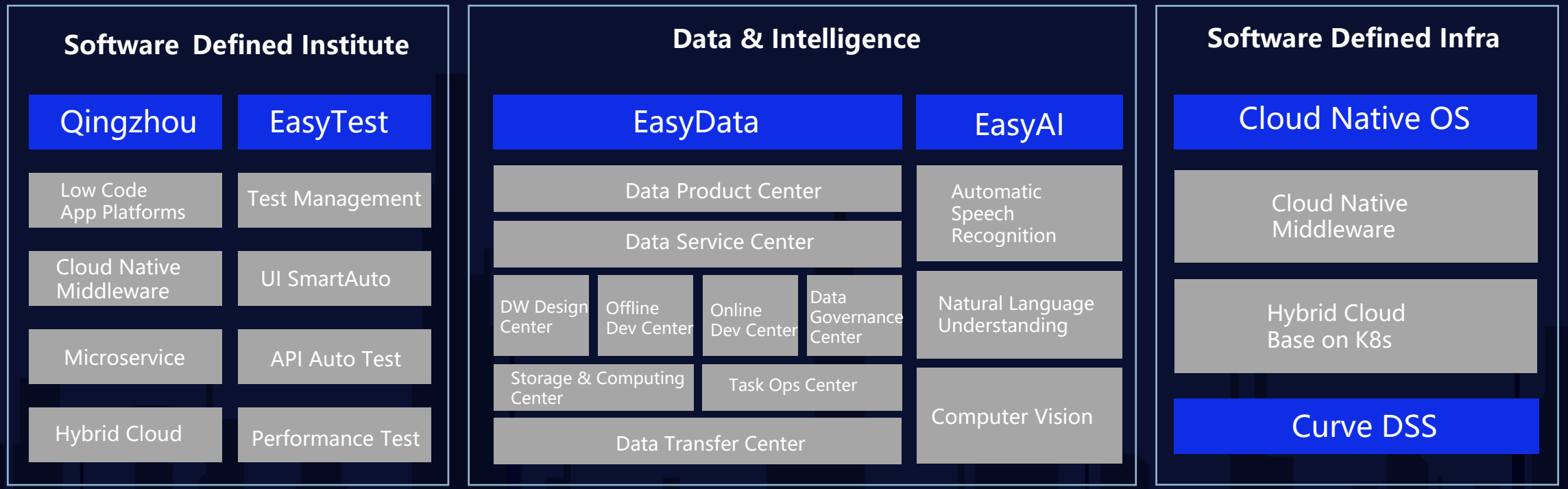
sodacon 2020

XiaoCui Lee

Senior System Development Engineer
NetEase Digitasail

Participated in the design and development of NetEase‘s self-developed distributed system—Nefs and Curve. Now is the project manager of Curve.

NETEASE DIGITALSAIL ENTERPRISE-DIGITAL-TRANSFORMATION SOLUTION ARCHITECTURE



OVERVIEW

Curve is a high performance, highly available, and highly reliable distributed storage system

- High performance, low latency
- Support storage scenarios: block storage, object storage, cloud native database, EC, etc
- Currently implemented high-performance block storage, docking openstack and k8s

  Service online nearly two years
- Open source
  - github homepage：https://opencurve.github.io/
  - github repository：https://github.com/opencurve/curve/

INTRODUCTION

- OVERALL DESIGN

  Basic structure | Data organization | Topology

- FEATURE

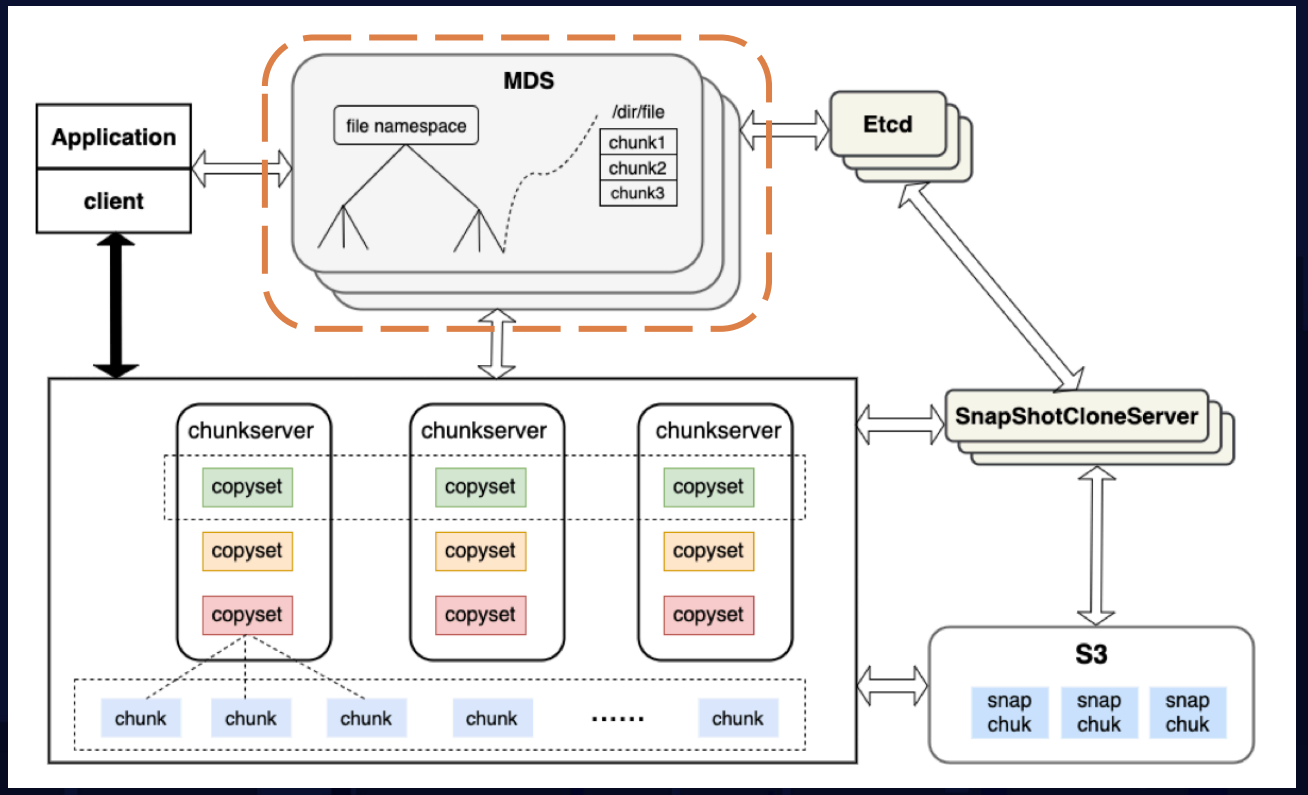
  High performance | High availability | Autonomy | Convenient maintenance

sodacon 2020
DATA CONNECTED

OVERALL DESIGN

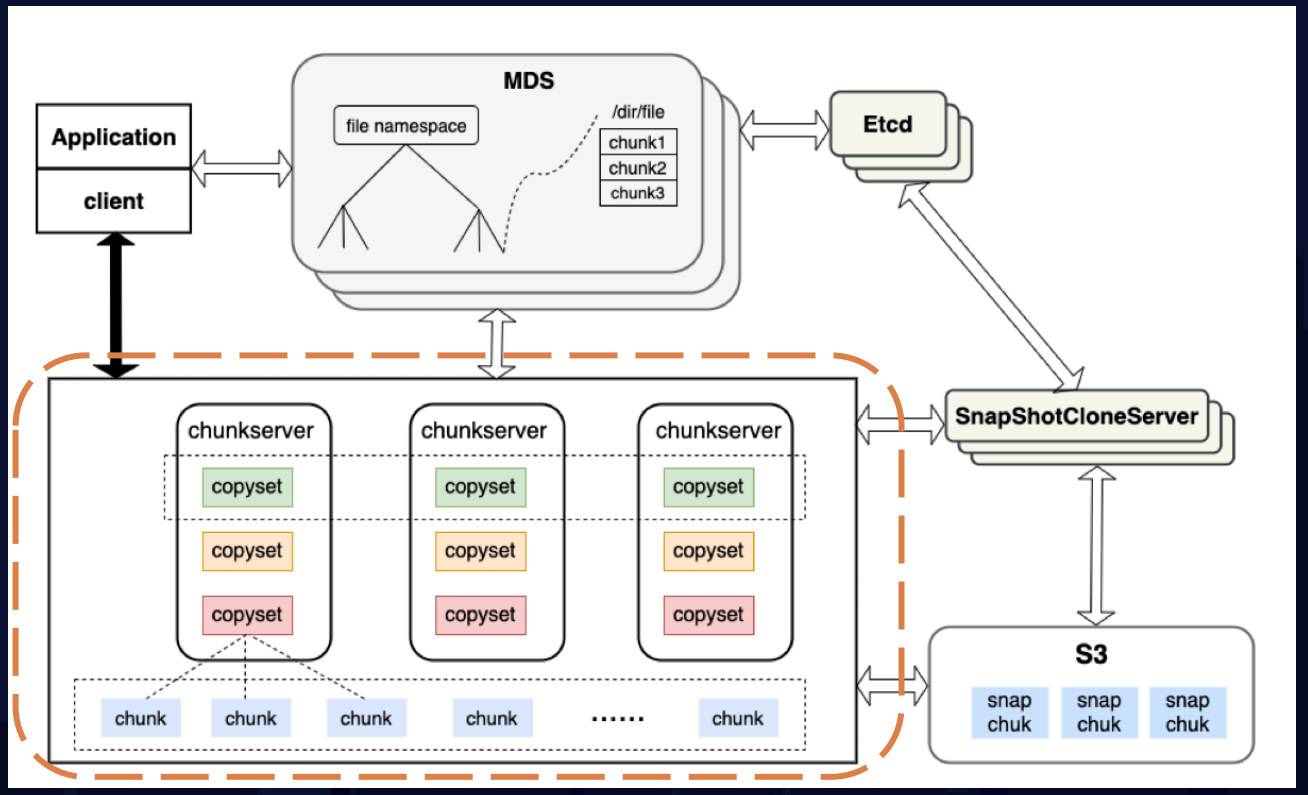
BASIC STRUCTURE

- MDS
  - Manage metadata
  - Collect cluster status information and schedule automatically



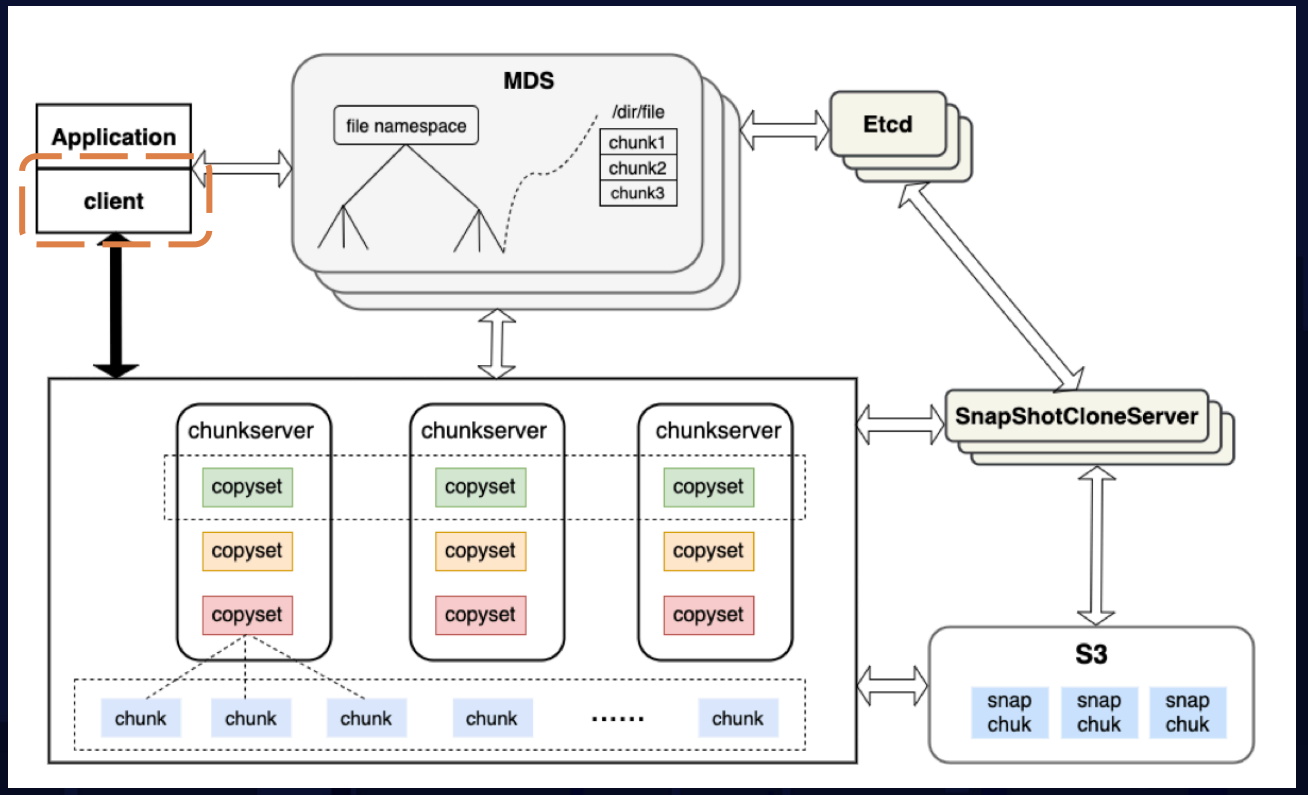
BASIC STRUCTURE

- Chunkserver
  - store data
  - data consistency



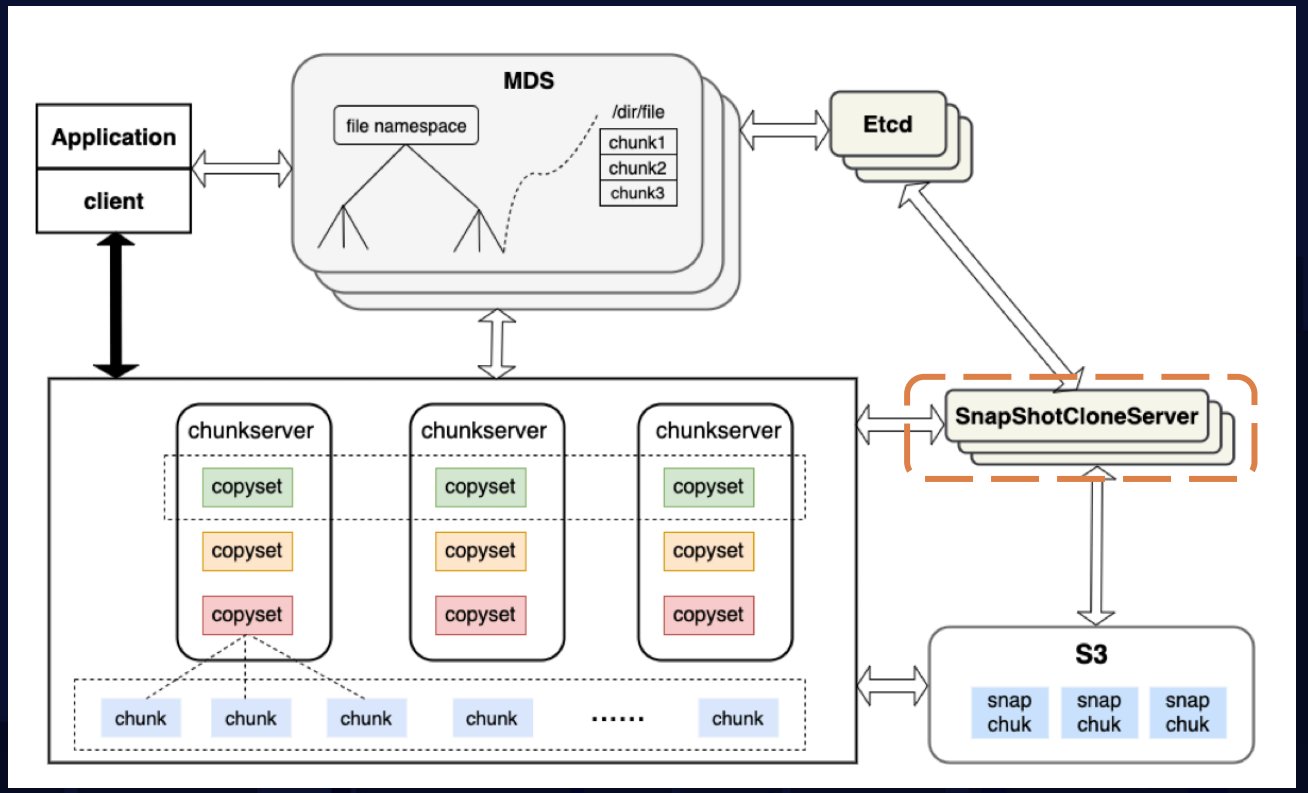
BASIC STRUCTURE

- Client
  - add / delete / modify / query metadata
  - add / delete / modify / query data



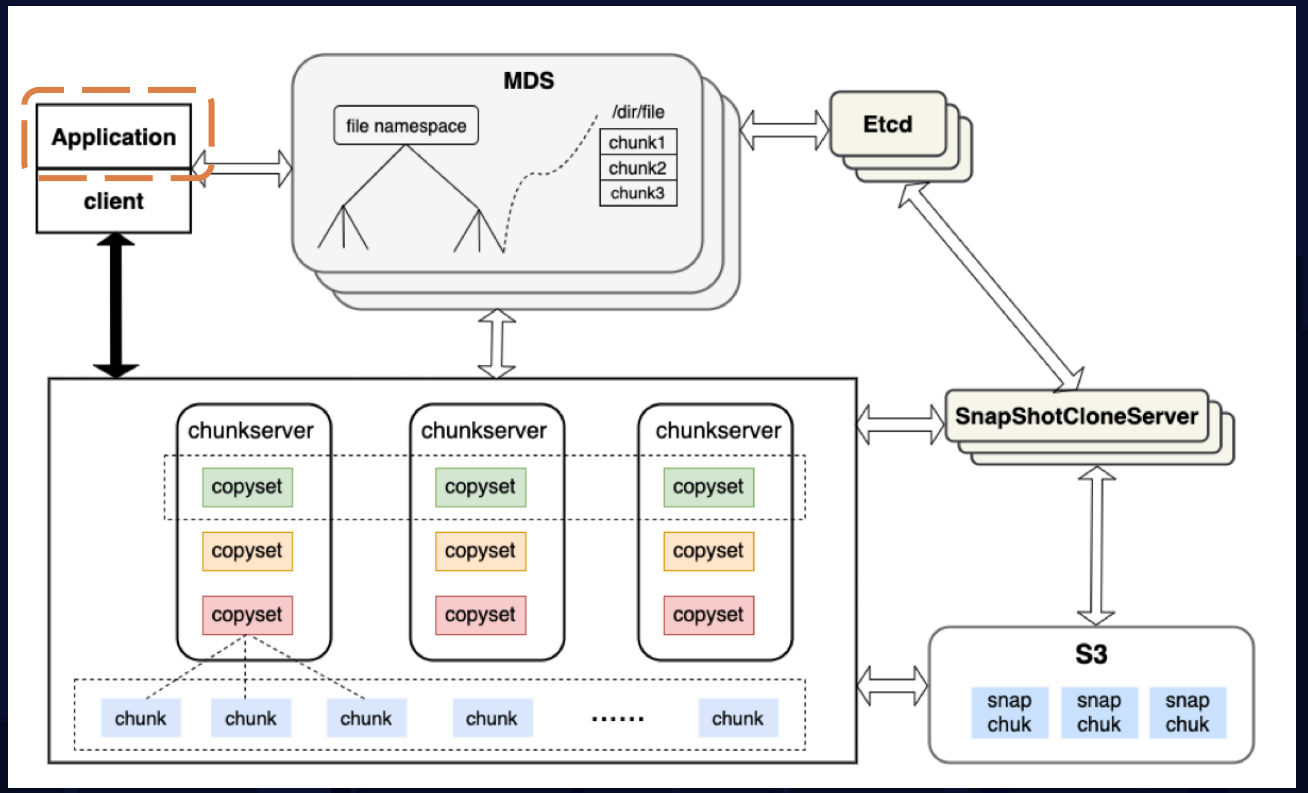
BASIC STRUCTURE

- Snapshotcloneserver
  - independent of core services
  - snapshot saved to object storage that supports the S3 interface
  - Asynchronous / incremental snapshot
  - clone from snapshot / mirror (lazy / no-lazy)
  - roll back from the snapshot



DATA ORGANIZATION

- Server
  - availability / reliability
  - scalability / load balancing
  - provide an undifferentiated file flow
- Applicaton
  - block / object / EC, etc
  - perceiving concrete formats

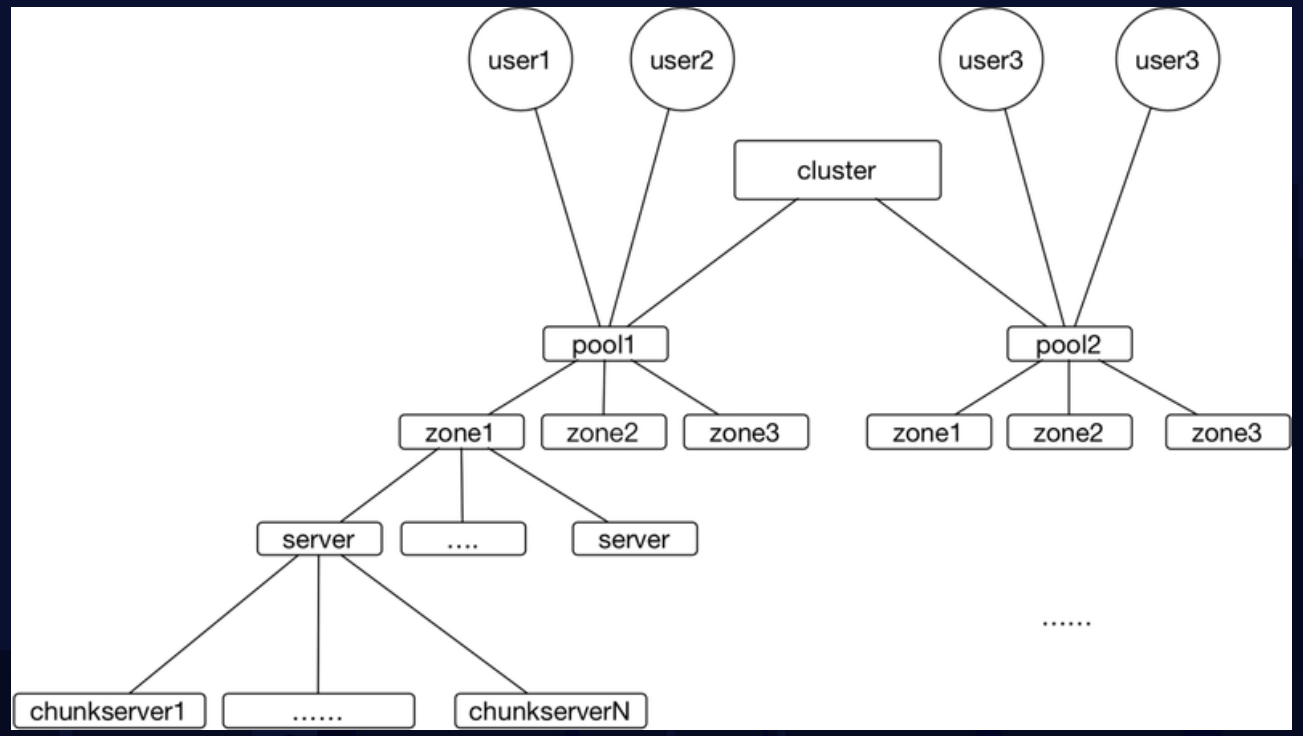


support different applications with different file types

(PageFile/AppendFile/AppendECFile)

TOPOLOGY

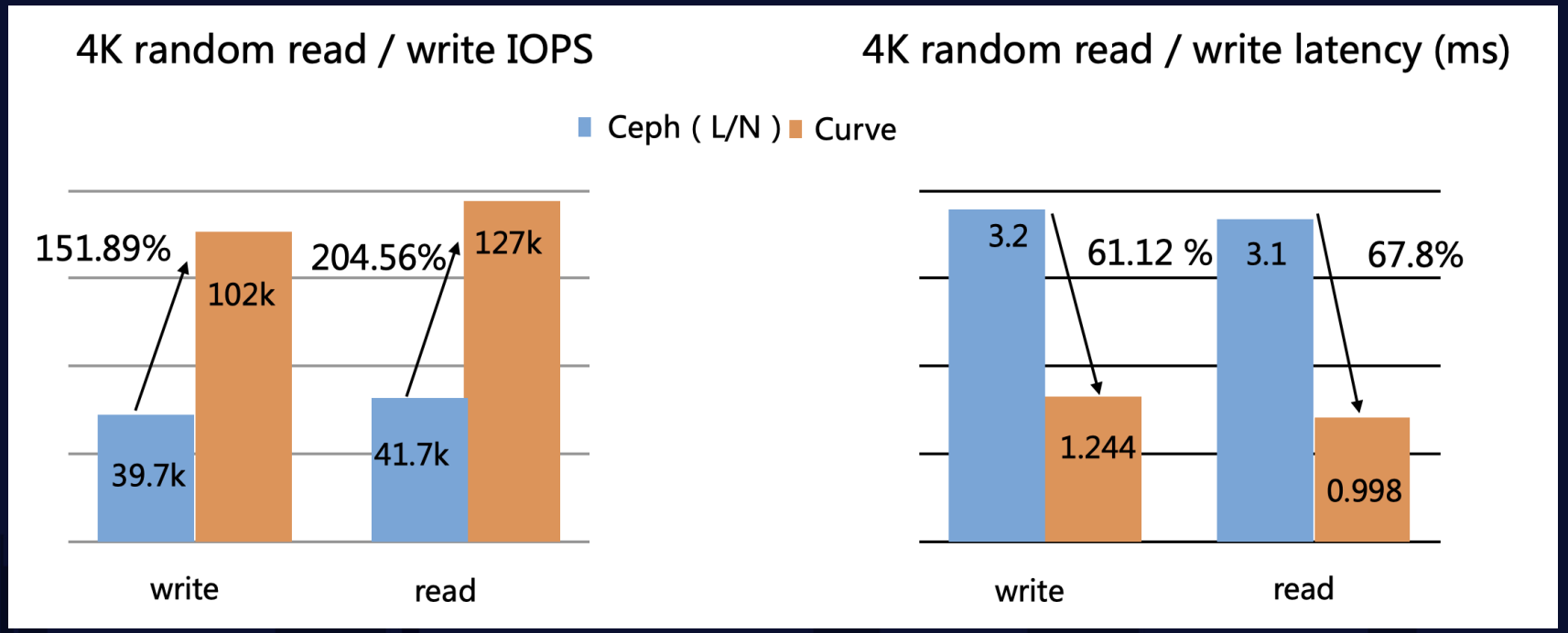
- Manage and organize machines
- Software: chunkserver
- Physical machine : server
- Failure domain : zone
- Physical pool : pool



sodacon 2020
DATA CONNECTED

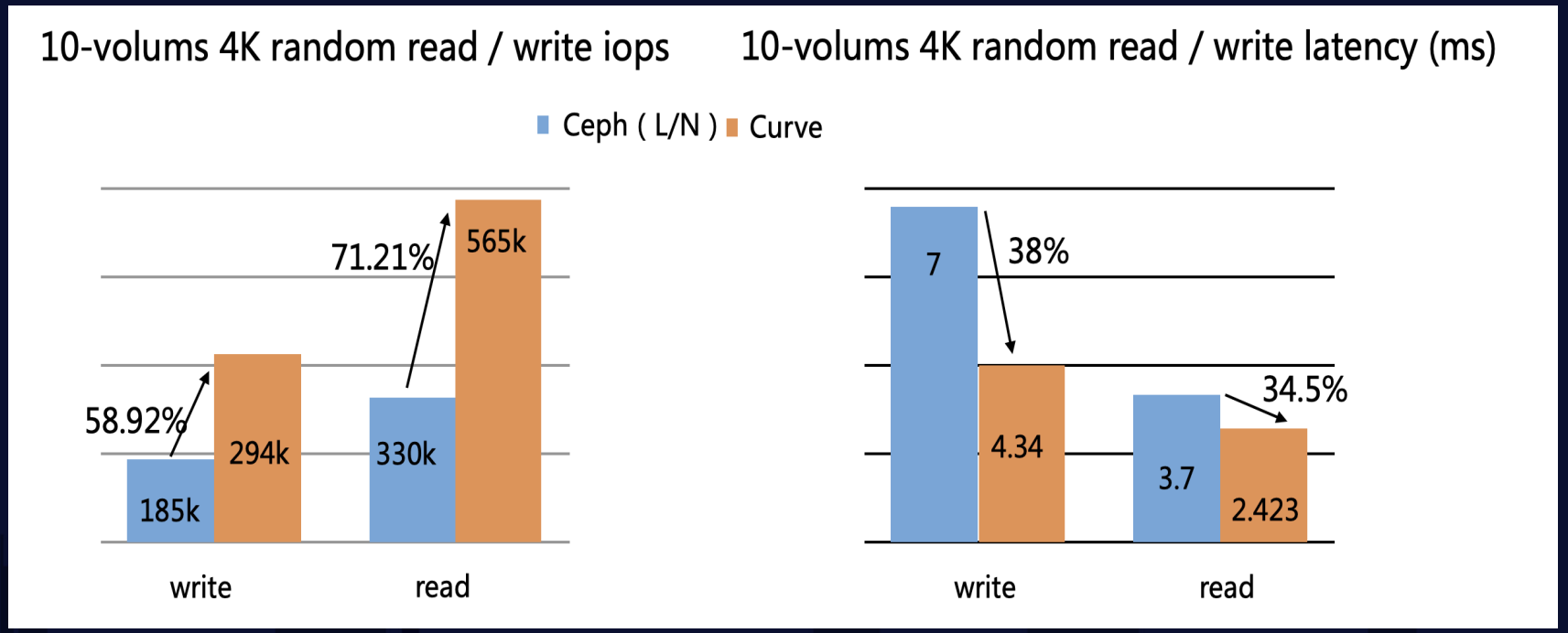
FEATURE

HIGH PERFORMANCE



Test environment: 6 servers each with 20 SATA SSDs
cpu: E5-2660 v4
memory: 256G
3 Replica scenarios

HIGH PERFORMANCE



Test environment: 6 servers each with 20 SATA SSDs
cpu: E5-2660 v4
memory: 256G
3 Replica scenarios

HIGH AVAILABILITY

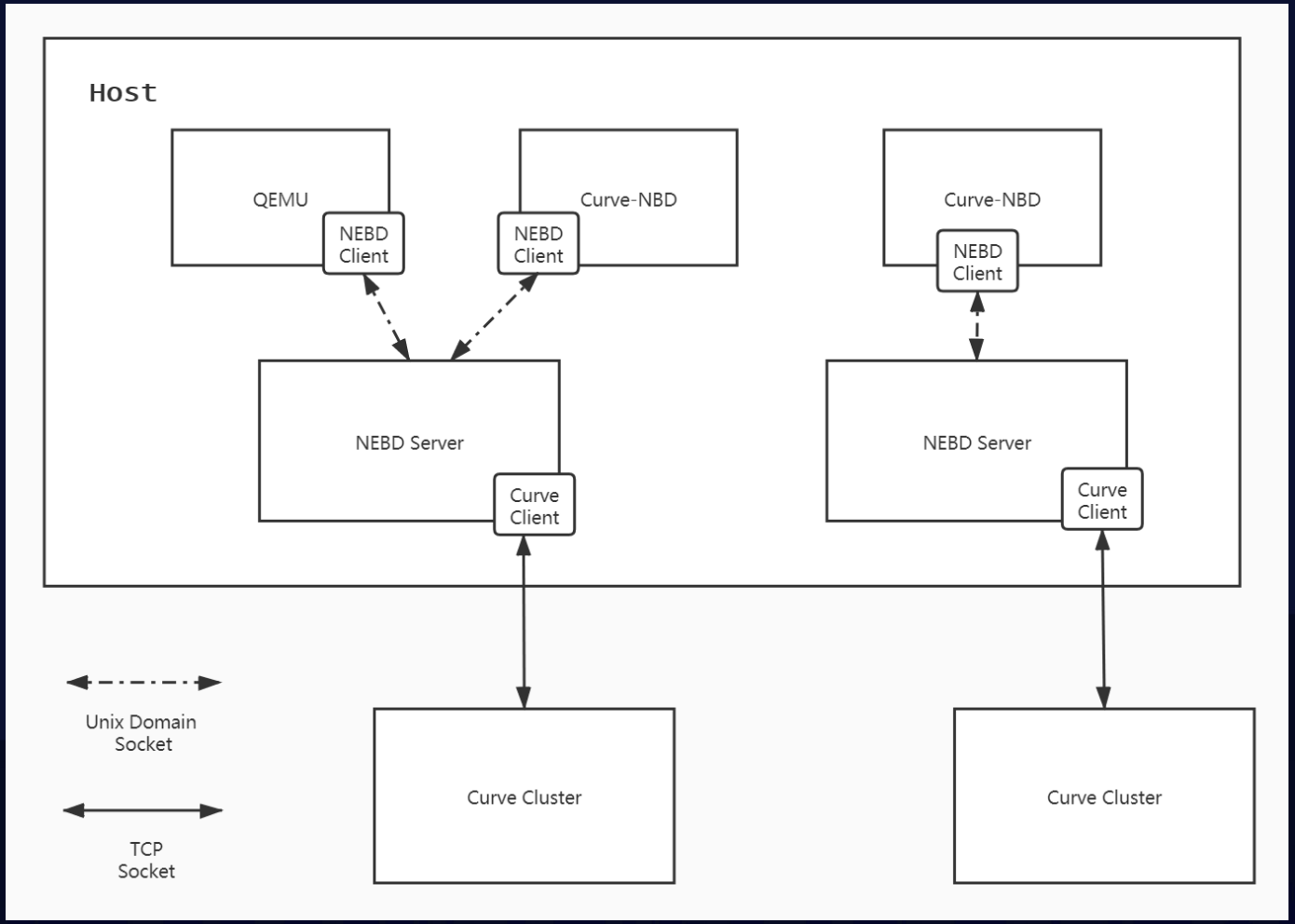
- The core component supports multi-instance deployment allowing partial instance exceptions
  - MDS, Snapshotcloneserver elect leader instance with ETCD
- Chunkserver achieve high availability with raft, 2N + 1 replica allows for N replica exceptions

AUTONOMY

- Automatic failure recovery
  - short recovery time
  - precise flow control with little impact on IO
- Automatic load and resource balancing
  - leader / copyset / scatter-width

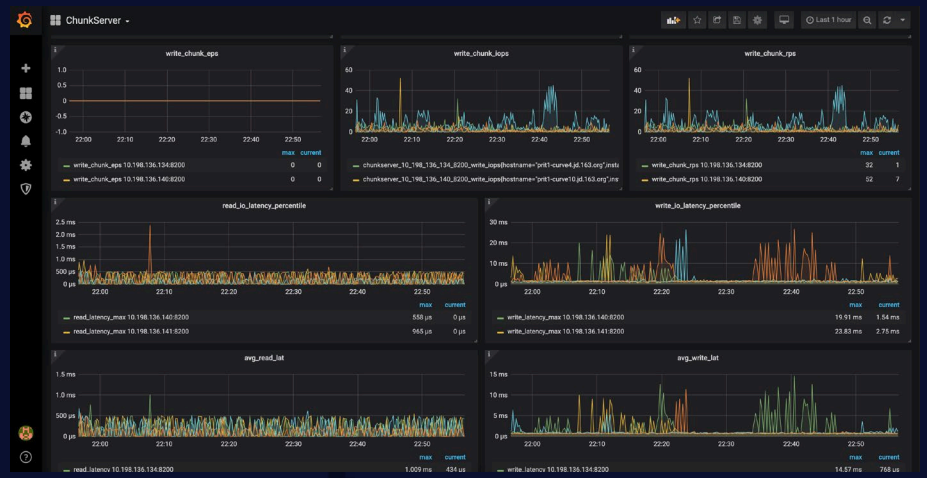
CONVENIENT MAINTENANCE

- Upgrade with second effect
  - client uses CS architecture
  - NEBD client : docking upper service
  - NEBD server : accept the request
    handle with curve-client
  - upgrade just restart the NEBD server

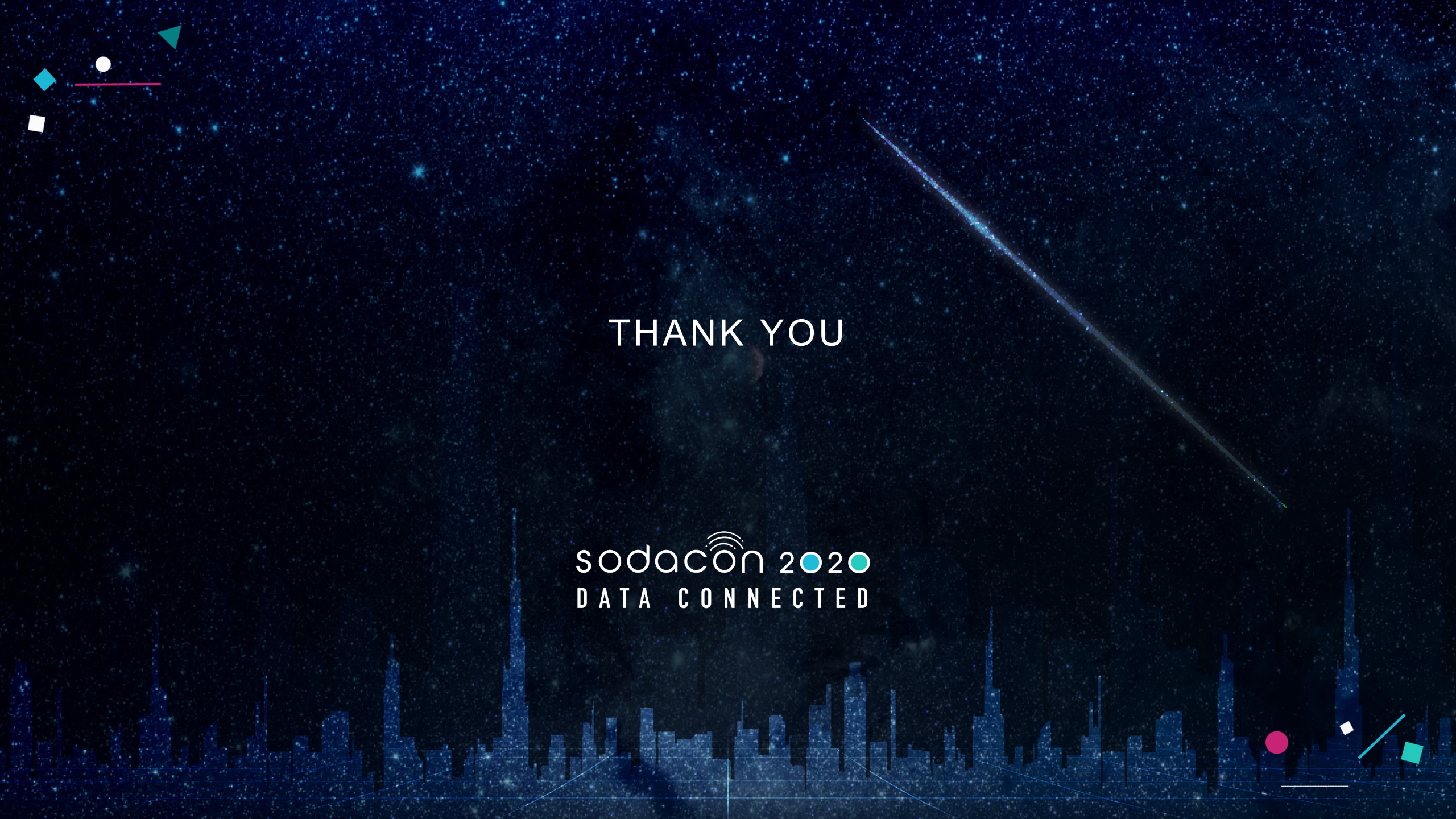


CONVENIENT MAINTENANCE

- Metric
  - collect with prometheus
  - visualization with Grafana
- Cluster status query tool
  - Curve_ops_tool
- Automated deployment tool
  - Develop with ansible
  - one-click deployment , one-click upgrade




```
Usage: curve_ops_tool [Command] [OPTIONS...]
COMMANDS:
space : show curve all disk type space, include total space and used space
status : show the total status of the cluster
chunkserver-status : show the chunkserver online status
mds-status : show the mds status
client-status : show the client status
etcd-status : show the etcd status
snapshot-clone-status : show the snapshot clone server status
copysets-status : check the health state of all copysets
chunkserver-list : show curve chunkserver-list, list all chunkserver infomation
get : show the file info and the actual space of file
list : list the file info of files in the directory
seginfo : list the segments info of the file
delete : delete the file, to force delete, should specify the --forcedelete=true
clean-recycle : clean the RecycleBin
create : create file, file length unit is GB
chunk-location : query the location of the chunk corresponding to the offset
check-consistency : check the consistency of three copies
remove-peer : remove the peer from the copyset
transfer-leader : transfer the leader of the copyset to the peer
reset-peer : reset the configuration of copyset, only reset to one peer is supported
check-chunkserver : check the health state of the chunkserver
check-copyset : check the health state of one copyset
check-server : check the health state of the server
check-operator : check the operators
rapid-leader-schedule: rapid leader schedule in cluster in logicalpool
```


THANK YOU
sodacon 2020
DATA CONNECTED