# Current practical experience with the distributed cloud data services based on iRDOS

George Jaroslav Kremenek

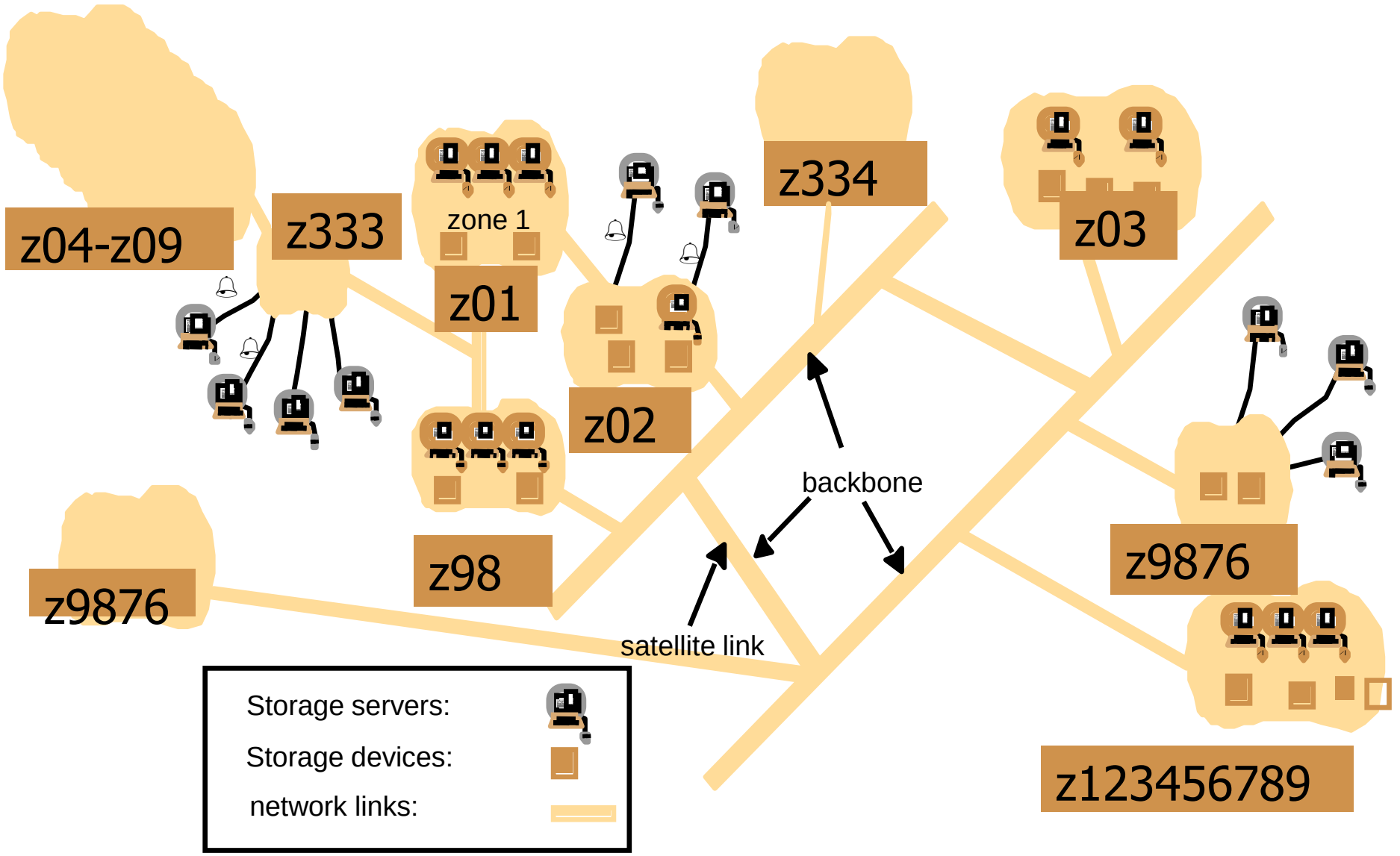
g.j.kremenek@gmail.com

George.Kremenek@cc.in2p3.fr

# Data explosion and iRODS (Swiss knife)

- We are currently witnessing data explosion and exponential data growth. iRODS is storing real PBs, in the near future EBs and hundreds or thousands millions of data sets. Classic file systems were not created to effectively manage thousands of millions data items. Inode space is usually limited. Storing very large data sets on disks is very costly (rotation + heat + cooling).

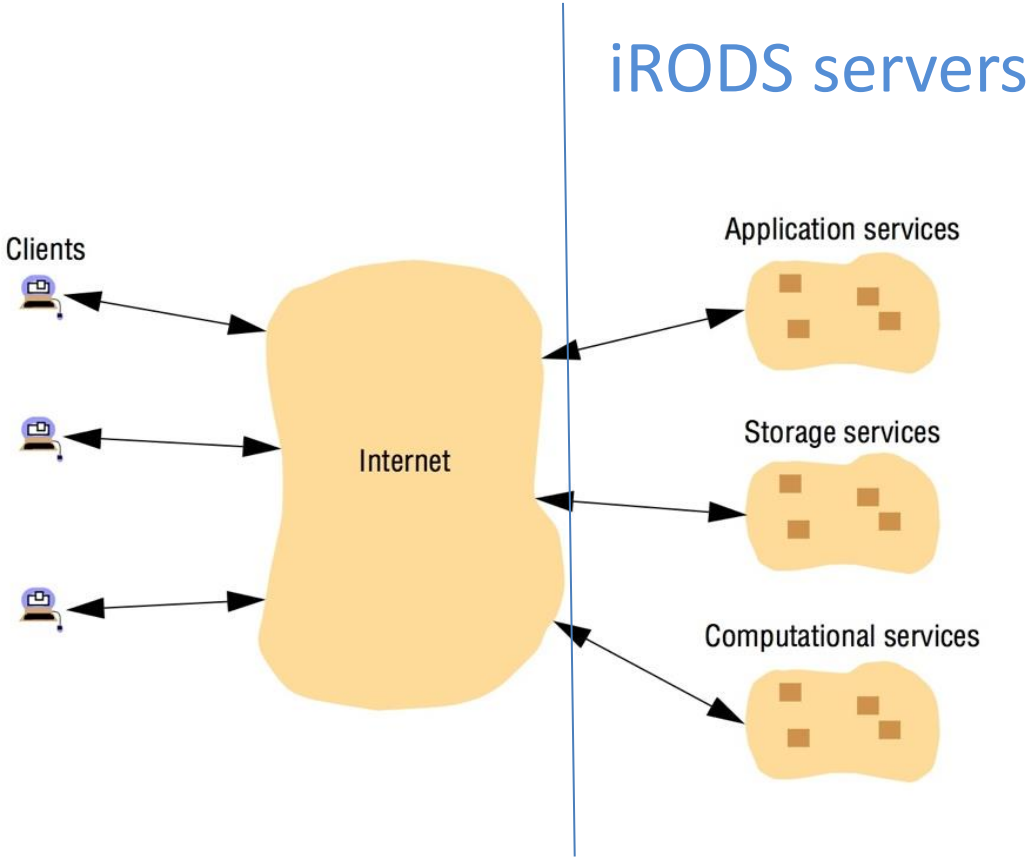
# iRODS Data Cloud Model (zones)



# iRODS Single Zone Model

http://opesol.org/cern.html (test it now)

- DBs
- iChat
- Data nodes
- Disks
- Tapes
- Clients: many
- interfaces, WEB, GSI,
- REST, GridFTP, etc.




- iR client contacts any iR server, this iR node contacts iChat, iChat contacts a RDBMS, than some other iR node contacts iR client

# Advantages and Disadvantages

- Disadvantages
  - Each zone has exactly one central iChat = RDBMS
- Advantages
  - Very flexible and extensible(ZTBs, RDBMS limit for # of items, TARs). Includes horizontal and vertical models.
  - RDBSM = standard SQL, it is very flexible and evolving
  - All current and future data management models can be incorporated (CASTOR, Lustre, LTO-FS, holographic cubes, etc.).
  - Can include other storage systems in its cloud model (GridFTP, XROOT, HSMs, HPSS, Posix, NTFS, RDBSM blobs, etc.)

# Data Management

- The long term data management is a key issue. It should ensure the availability and continuity of access for ever. Scientific collaborations are usually geographically dispersed, which requires the ability to share, distribute and manage world wide efficiently and securely. The media, hardware and software storage systems used can differ greatly from one scientific center to another.

# Heterogonous

- Different scientific centers will always use heterogonous HW and SW. They will add to and modify their respective HW and SW with the continuing IT system evolution. This induces continuous physical and logical data migration. Technological development and innovation in SW may involve changes in the naming or data access protocols while backward compatibility must be maintained.

# Middleware to the rescue

- Such environments can take advantage of middleware for the management and distribution of data in a heterogeneous environment, including virtualizing storage, that is to say, hiding the complexity and diversity of systems underlying storage while federating data access.
- Virtual distributed hierarchical storage system, data grids or data clouds require using and re-using existing underlying storage systems. Creating completely new vertically integrated systems is out of the question.

# iRODS approach

- iRODS based solutions can take advantage of existing HSM like IBM's HPSS and TSM, SGI DMF, ORACLE (SUN) SAM QFS or emerging cloud storage system like Amazon S3, Google, Microsoft Azure , Hadoop and other. A good middleware distributed cloud service for very large data sets should work with all main existing such system and be extensible enough to support main future systems.

# History and present

- iRODS (integrated Rule based Data System) is being developed in USA for over 20 years.
- http://www.irods.org
- Noncommercial, commercial support exist.
- User groups.

# iRODS - Open Source Code

- iRODS is HW and company agnostic and the users have access to all the source code. Migration from one storage resource to another (new) one is just one iRODS command regardless of the data size or number of data objects.
- But what makes iRODS particularly attractive is its rules engine that has no equivalent among its competitors. The rules engine allows complex tasks at data management.

# iRules and Data Management Policies

- These user defined policies perform remote data management at the server side: for example, when data is stored in iRODS, background tasks can be triggered automatically on the server side such as replication across multiple sites, data integrity checks, post-treatment on them (metadata extraction, etc.) without specific action on the client side. So, the management policy data is virtualized. This virtualization ensures strict rules set by users, regardless of location data or application that accesses iRODS.

# LT-FS and iOLTFS

- iRODS like systems can deliver full vertical data storage stack including complex tape system management using the existing modern open and standard LT-FS technology. The LT-FS (http://en.wikipedia.org/wiki/Linear_Tape_File_System) exists on all modern tape drives and tape libraries.

# iRODS sites

- I will talk today about sever sites which have chosen a data grid system based on the iRODS (Rule-Oriented Data management) system. IRODS provides a rule-based system management approach which makes data replication much easier and provides extra data protection. Unlike the metadata provided by traditional file systems, the metadata system of iRODS is comprehensive and extensible by user and allows users to customize their own application level metadata. Users can then query the metadata to find and track data.

# IRODS at CC IN2P3

- The Computing Center of L'Institut national de physique nucléaire et de physique des particules (IN2P3, CCIN2P3) offers iRODS service IN2P3 since 2008. This service is open to all. Currently it is used by 34 groups in the fields of particle physics, nuclear physics, astroparticle physics and astrophysics. The CC-IN2P3 provides hosting to the central catalog of iRODS Grille a new French data  service. The iRODS service at CC IN2P3 has its own data and disk servers and it is interfaced with the IBM HPSS mass storage (tapes) currently managing over 8PBs of data.  The iRODS services are federated with other iRODS services elsewhere for example at SLAC USA.

# iRODS at BnF

- The **Bibliothèque nationale de France** (Bnf) is using iRODS together with open (closed) SAM QFS to store hundreds of million books in its Long-term data preservation. BnF is using iRODS to provide a distributed private data cloud where multiple replicas of data sets are kept at primary BnF site in Paris and secondary site about 40 km From Paris. BnF created a tool to implement its policies for digital preservation SPAR System (Distributed Archiving and Preservation), launched in May 2010, and is continually updated with new collections and feature. BnF employs SPAR and Gallicca for the WEB interface to the distributed private data cloud in iRODS.

# iRODS at NKP and NDK

- **The NKP and NDK site is a EU project (Czech National Library, Czech National Digital Library).** I have helped to implement iRODS together with Fedora Commons and other tools at NKP in Prague Czech republic as a base for the EU funded digital library project. The system is now in a full production. It is using IBMS GPFS and TSM as a base layer for its HSM.
- The system stores over 300 million data objects. Its data comes nonstop from scanning paper books, electronic data input from Born Digital documents, constant WEB archiving of the “.cz” domain and from all Czech TV and radio broadcasts among others.

# Open test sites and examples

- http://opesol.org/login.html
- http://opesol.org/login2.html
- http://opesol.org/login3.html
- virtual tape library (VMware Player WIN7)

# Service summary

| | |
|---|---|
| Status: | Production |
| Number of users (current, target): | 22/NL, 33/NL, 44/NL (NL = No Upper Limit) |
| Default and Maximum quota: | N/A, N/A, N/A, |
| Linux/Mac/Win user ratio: | 98% / 1% / 1% |
| Desktop clients/Mobile Clients/Web access ratio: | 79% / 1% / 20% |
| Technology: | Own Cloud with DB using iRODS |
| Target communities: | researchers and users in many fields |
| Integration in your current environment (examples): | migration between the sync service and the batch facilities is transparent |
| Risk factors: | Very low |
| Most important functionality: | Heterogeneous, extensible, HW, SW a company agnostic |
| Missing functionality (if any): | None? Extensible by users. |

# User feedback

- Examples
  - Love it!
  - Loathe it!
  - Live with it!
  - User always want more functionality (Swiss knife), (versioning, etc.) and more performance

# Tests, demos, feedback, questions

- **http://opesol.org/cern.html** (test it for yourself)
  - Offline demo available: my laptop, Wmware, iRODS with LTO6 robot (IBM 3584 Ultra Scalable Library, IBM 3580 Ultrium6 drives, LTFS, LTO6)

# •Questions?