Database Competence Centre

openlab Major Review Meeting 2010

25th January 2011

Zbigniew Baranowski
Andrei Dumitru
Carlos Garcia Fernandez
Anton Topurov
Dawid Wojcik

CERN
openlab

Outline

- Oracle VM at CERN
- WebLogic Server on JRockit-VE
- Oracle JRockit Mission Control
- Oracle Complex Event Processing NEW!
- Enterprise Manager
- GoldenGate 11g
- Replication Technologies review
- ASM-based Cluster File System 11.2 tests
- Outreach

Oracle VM at CERN

Status January 2010

- **Only some** paravirtualized **test** VM installed manually on single machines via XEN
- **Many** blocking **problems** integrating with CERN's infrastructure to be solved
- Different **deployment strategies** were being explored

Context:
The blocking problems were due to the complex CERN infrastructure, that obliged us to follow strict network rules and integrate Oracle VM as any other Linux distribution in our provisioning system

Status January 2011

- **3** Clusters of OracleVM Servers running **53** VMs fully hardware virtualized
- **Multi-layer architecture**: Storage, Server, Virtual, Logic/database (manager), Monitor
- Oracle clusters **(RACs) of 11g** installed in virtual machines
- **Quattor installation** of hosts and guests (RHES5, RHES4, SLC5)
- **OracleVM Manager** with **redundant configuration**, **OVM database on our production DB**

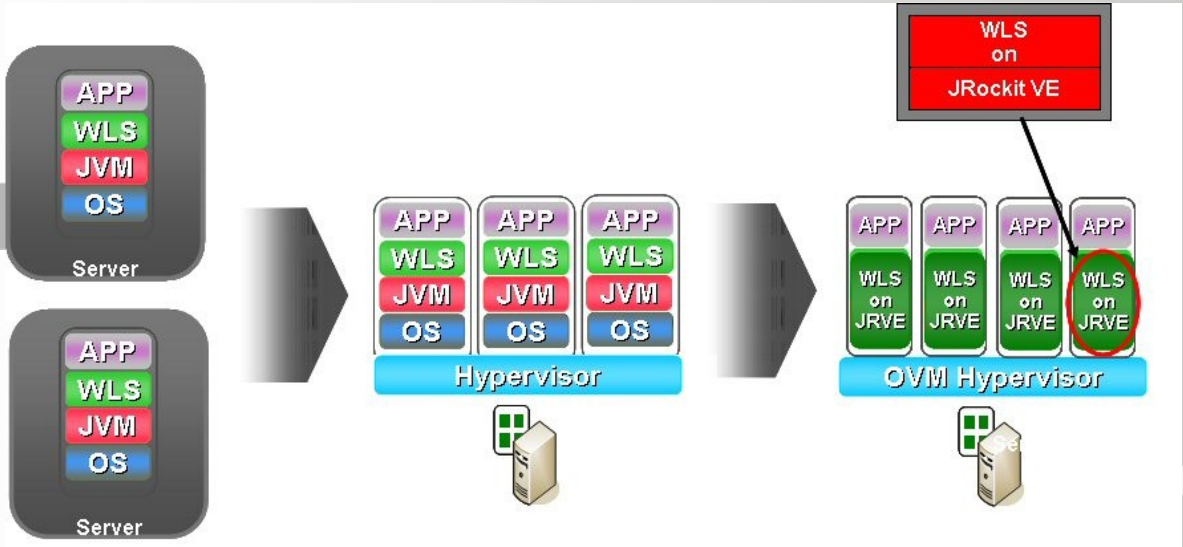
Status January 2011

- **Automatic creation** and **reinstallation** of VMs
- Multiple **Gigabit Ethernet** interfaces and fail-over **bonding** support on VM
- **High Availability** active:
  - Automatic cluster server master node migration
  - VM live migration
  - Power-cut tolerance tested
- **2 interfaces of operation integrated**:
  - Web
  - Command line

Status January 2011

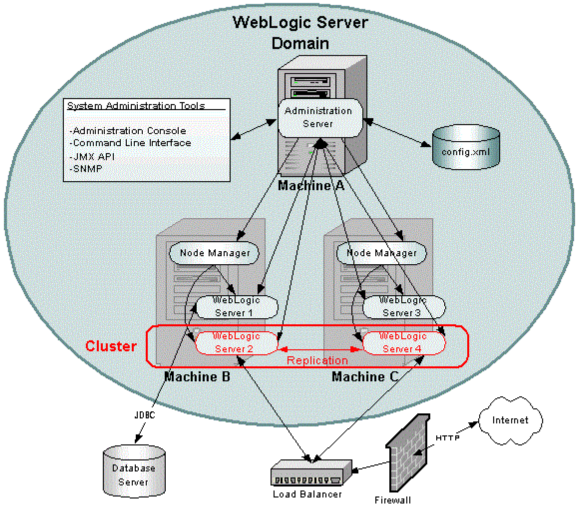
- **Service Level Status** Monitor integration
- **Firewall** and **OS-update** configured.
- Multiple **optimization** explored and implemented
  - VM fast search
  - Memory ballooning
- Success of multiple **OS** and **Database Performance** test, including database recovered from tape
- Used in **production** for databases and application servers

WebLogic Server on JRockit-VE (WLS)



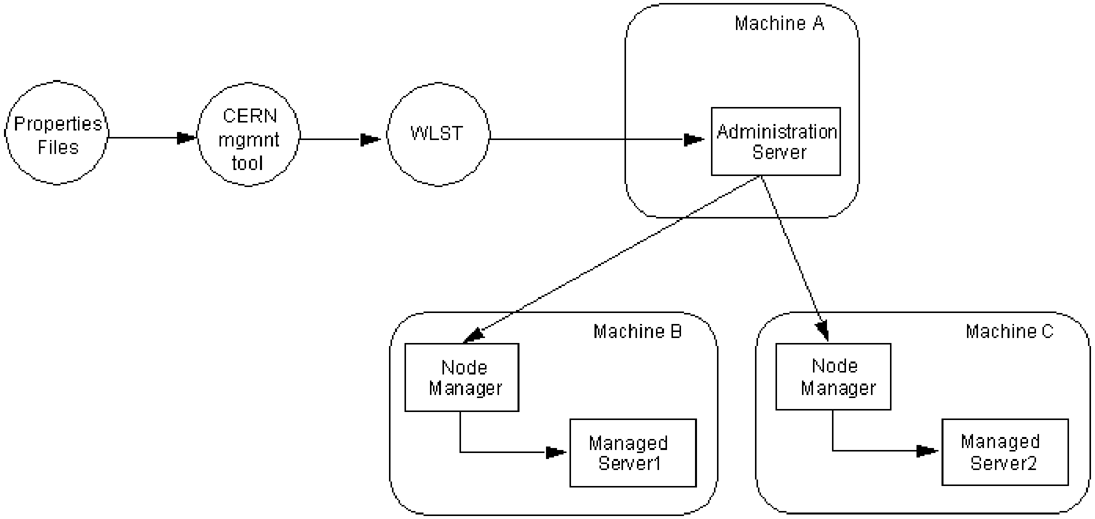
Ensuring repeatable deployment

- **Integrate** WLS management with **Syscontrol** (operations with domains)
- **Simplify domain configuration**:
  - Automate all processes with scripts
- **Recreate domain** when needed:
  - Repository of configurations



Implementation

- WebLogic provides **WLST** which is a scripting interface tool
- **Properties files** which represent domain configuration
- **Jython scripts** to manage WebLogic installations and perform configuration actions



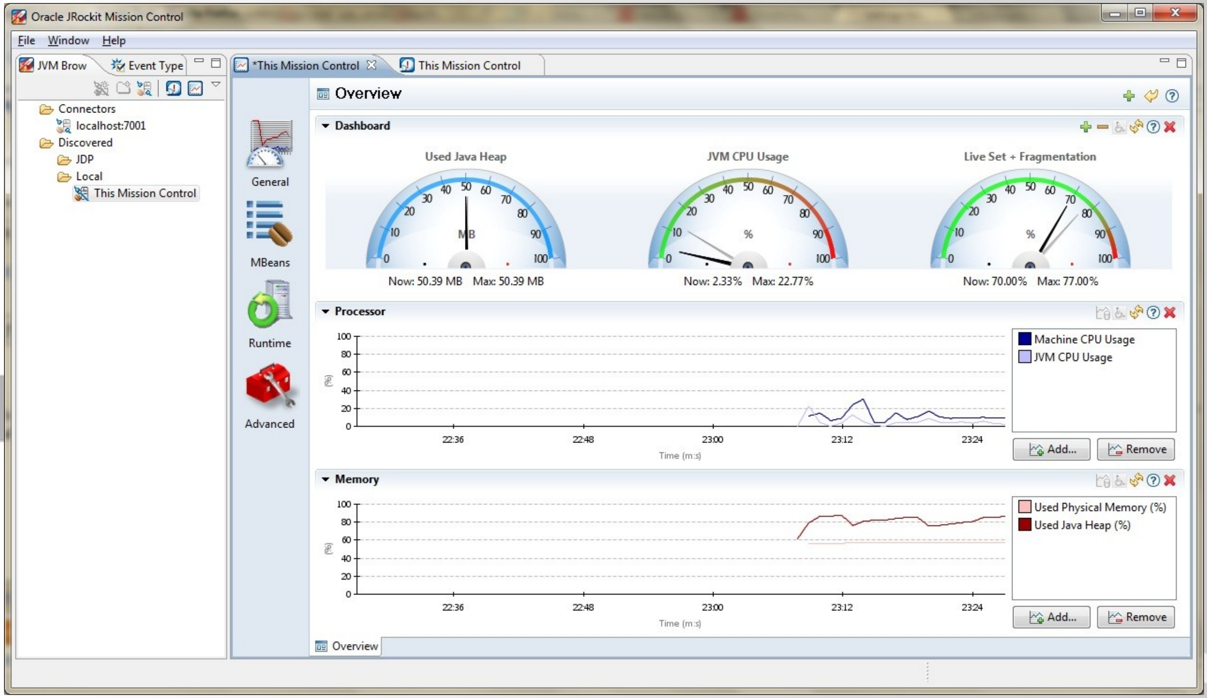
Differences physical vs virtual

- **Not many differences** thanks to WLST
- Some **commands** used for physical servers have just been **replaced** by their virtual counterparts
  - **Node manager** implemented on **Oracle VM manager**
- Validated with **both physical and virtual** servers in the same domain
  - Even if not a realistic deployment configuration

Operations

- Create domains
- For each managed server:
  - Create virtual machine
  - Set VM parameters
  - Inject required files by applications
- Start admin server
- Create, configure and assign servers into clusters
- Interface for developers to deploy applications

Oracle JRockit Mission Control



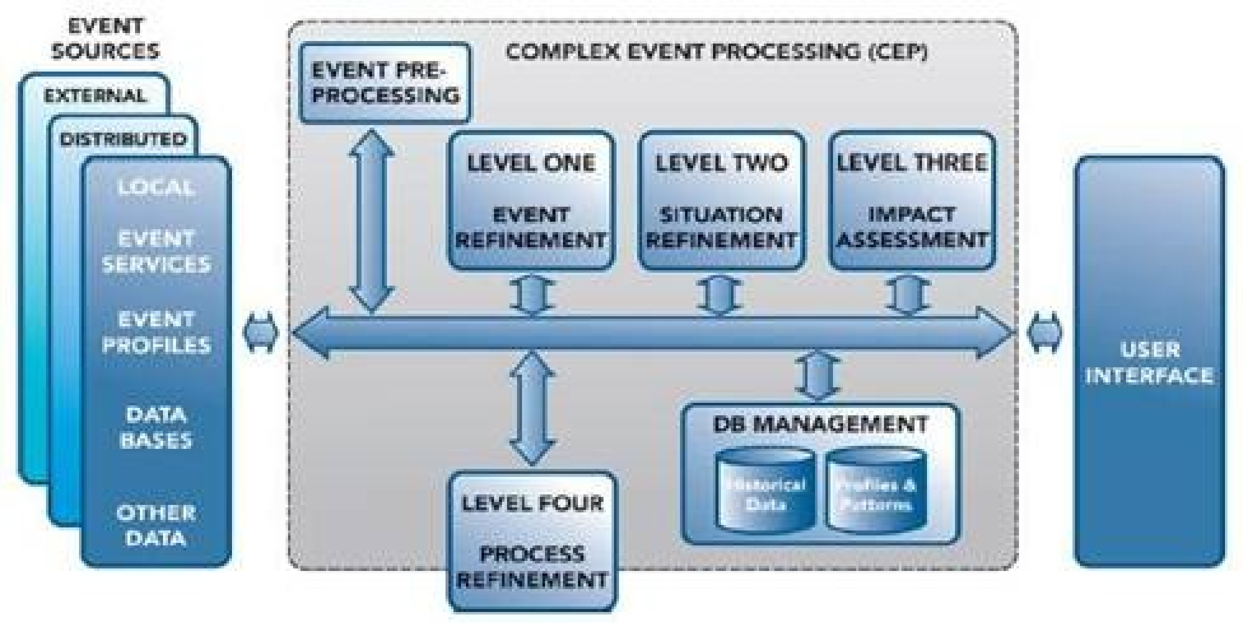
JRockit Mission Control

- Oracle JRockit Mission Control Client is a suite of tools:
  - **Monitor**
  - **Manage**
  - **Profile**
  - Without any performance overhead

- Presented by Oracle in the openlab framework to CERN's developers
- Helped to solved really difficult to debug application **memory leaks**

Oracle Complex Event Processing

CERN openlab



EVENT SOURCES

EXTERNAL

DISTRIBUTED

LOCAL

EVENT SERVICES

EVENT PROFILES

DATA BASES

OTHER DATA

COMPLEX EVENT PROCESSING (CEP)

EVENT PRE-PROCESSING

LEVEL ONE

EVENT REFINEMENT

LEVEL TWO

SITUATION REFINEMENT

LEVEL THREE

IMPACT ASSESSMENT

LEVEL FOUR

PROCESS REFINEMENT

DB MANAGEMENT

Historical Data

Profiles & Patterns

USER INTERFACE

Oracle CEP

- Oracle **Complex Event Processing** (CEP) is a complete solution for building applications to **filter**, **correlate** and **process events in real-time** so that downstream applications, service oriented architectures and event-driven architectures are driven by **true**, **real-time intelligence**.
- Possible application to Security and Network Traffic analysis

Enterprise Manager

Enterprise Manager 11g

- New functionalities to monitor Middleware
- CERN upgraded in September 2010
  - Early adopter
- Our configuration/workload highlighted unexpected memory issue
  - **Heap usage** is more than expected
  - Research is **ongoing** to spot the root cause.
  - openlab involvement very helpful escalating within development

WebLogic Monitoring

- **Discovery** of **WebLogic Domains** using *emcli (EM Command Line Interface)*
  - Discover domains automatically
  - Custom script to enable/disable the refresh domain job
  - More tests to be done on production environments

- **Monitoring Templates** application
  - Applied daily using *emcli*
  - Scripts adapted to use 11g

Integration with management tools

- Integration with **syscontrol**
  - Syscontrol to be **single point of truth**
  - Targets to be created based on content in syscontrol
    - **Add** RAC to EM Grid Control using only ***emcli***
    - **Not** based on **auto-discovery**, not straightforward
    - Work in progress

Next activities

- **Automatic grouping** of targets based on syscontrol
- More **tests** for **Weblogic** monitoring
- **Integration** with existing tools like ***State Management System*** *(SMS)* for a consistent view of Service Status during interventions
- Next Generation EM **beta tests**

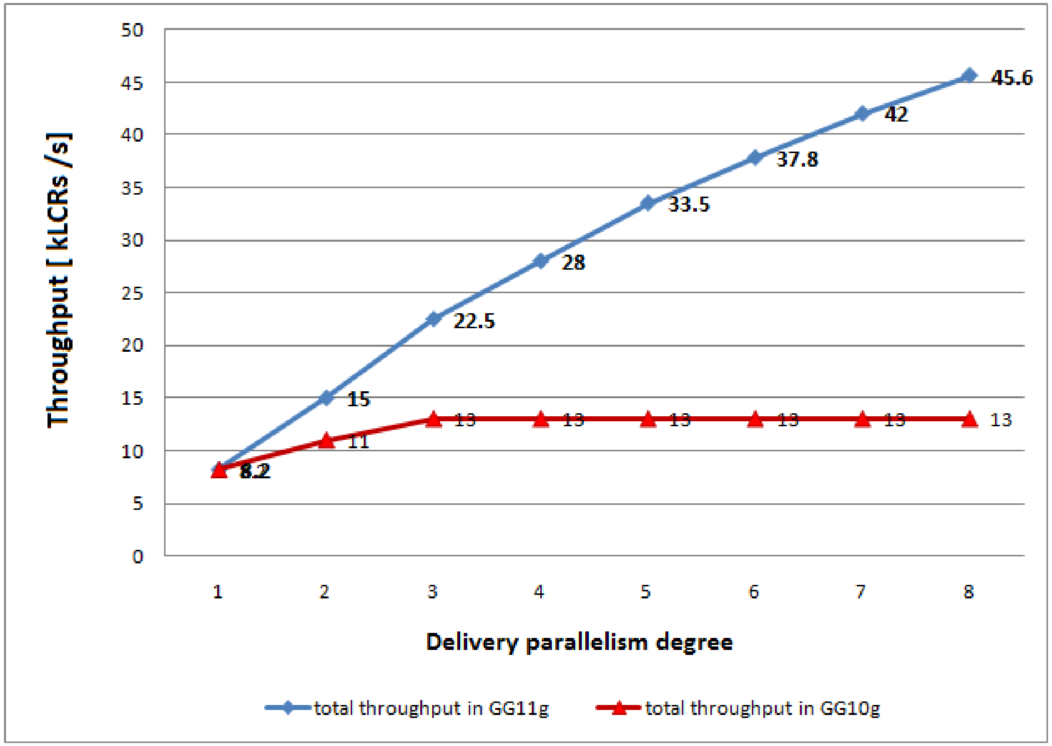
GoldenGate 11g

GoldenGate

- New replication technology from Oracle
- Provides real-time data integration across **heterogeneous environment**
- Long term plans for integration with Oracle Streams
  - To be **as fast and functional** as Streams11g
  - To be **as stable and flexible** as GoldenGate
- GoldenGate11g is the latest version released last September
- Tested on 10.2.0.5 and 11.2.0.2
- New Automatic Storage Management (ASM) parameters have been tested
  - No need to specify ASM access credentials
  - No performance impact

GoldenGate11g performance

- Delivery process is the **bottleneck** (as in 10g)
- Default delivery parallelism is not scalable
- One **delivery process per schema**
  - Improved performance – scalable throughput



Context:
Golden Gate can scale , but on a per schema basis.
This might be compelling in general, but not for CERN’s use cases.

GoldenGate11g - summary

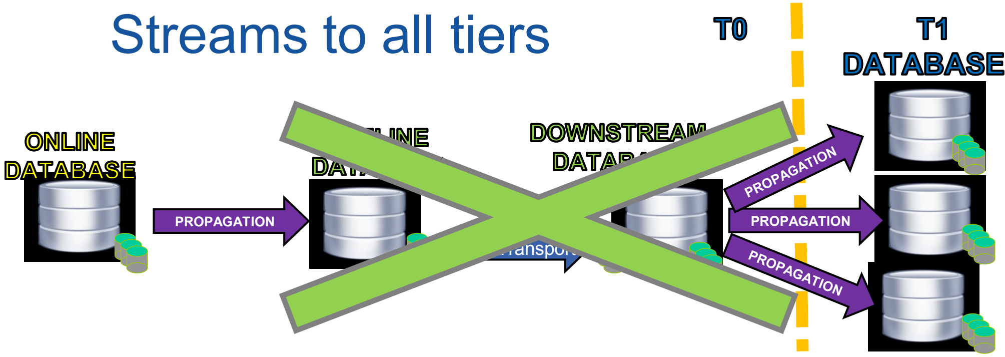
- **Stable and reliable** software for heterogeneous database replication
- **Improved performance** similar to Streams11g under **certain conditions** (workload generated by multiple users)
- Still some **problems** with handling of **data definition** modifications
  - GoldenGate is focused on pure data transfer
- **Monitoring** software is not available
- Currently cannot be used in combination with Oracle DataGuard – source database has to be read-write
- CERN feedback positively appreciated by Oracle

Replication technologies review

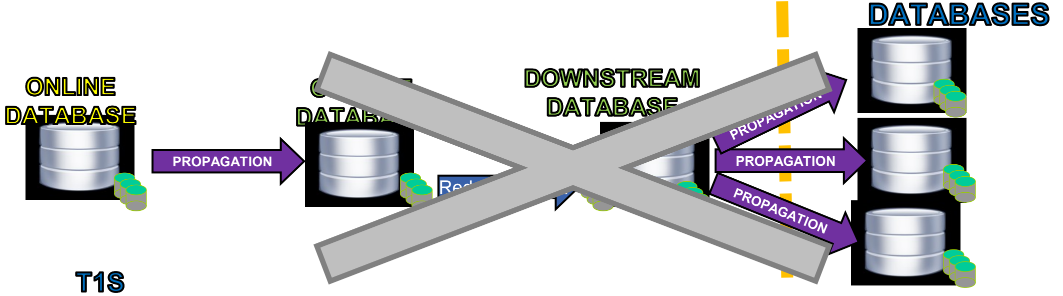
Replication technologies

- New releases of replication technologies has been reviewed as part of preparation for **migration** of databases to version **11g**
- Proposals of replication solutions were presented to **experiments and T1s DBAs** in collaboration with Oracle representative during last ***Distributed Database Workshop*** *@CERN* in November

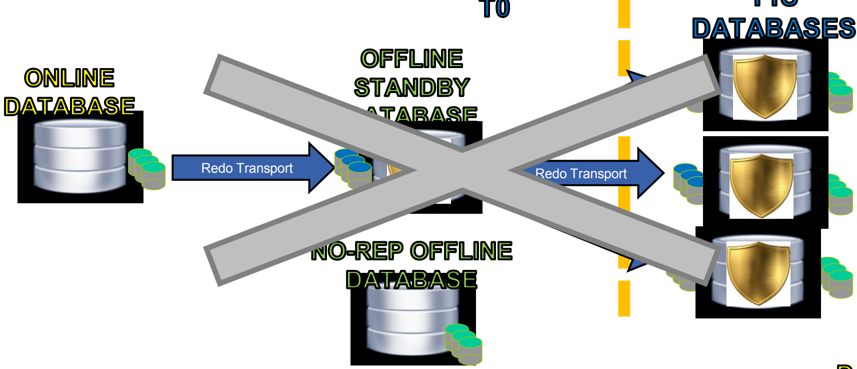
Streams to all tiers



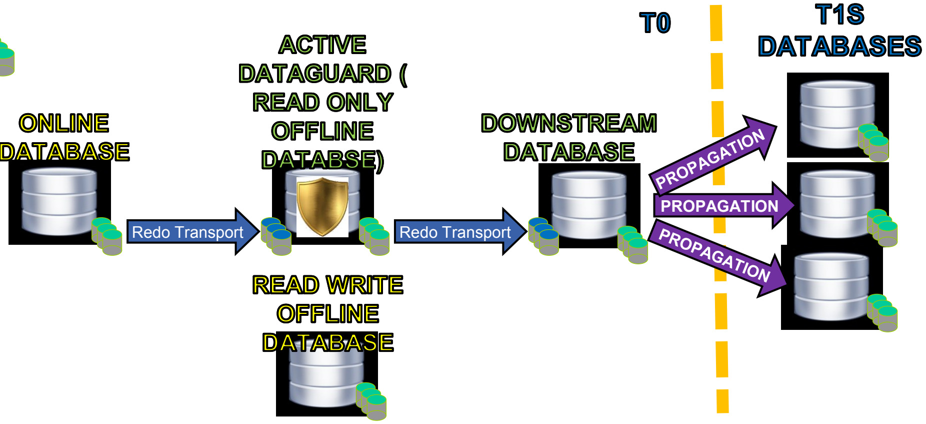
Goldengate to all tiers



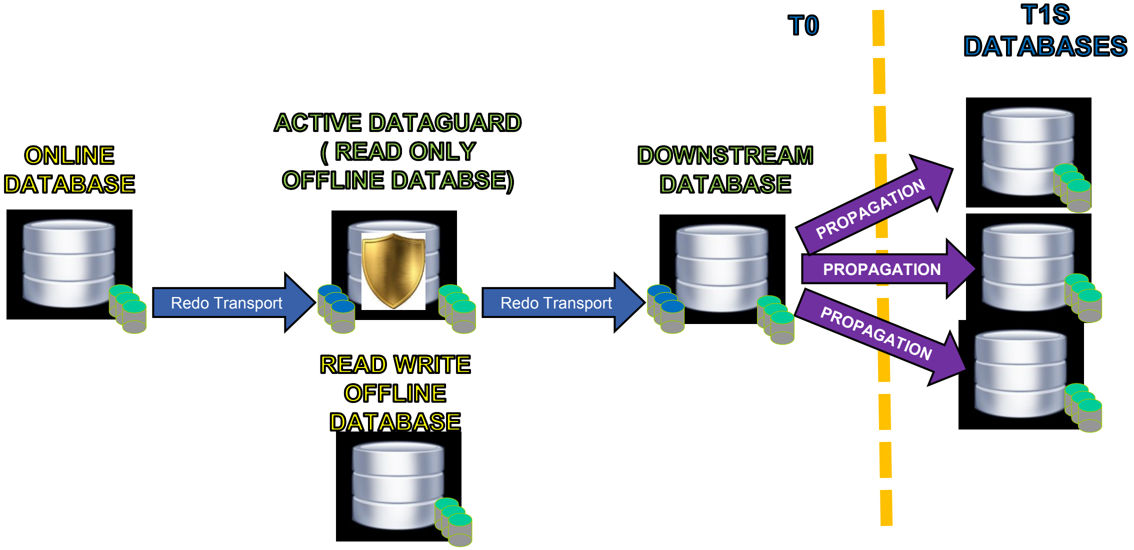
Active DataGuard to all tiers



Active DataGuard & Streams



ActiveDataGuard11g & Streams11g



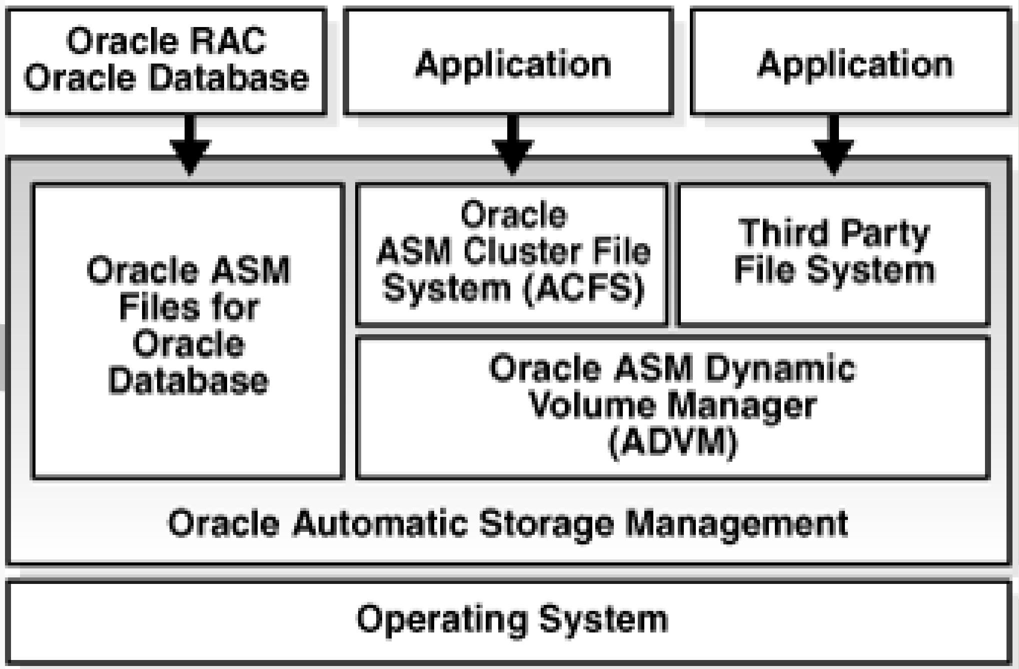
- ☺ **Fast and reliable** ONLINE-OFFLINE replication
- ☺ **Lower maintenance** effort (physical replication) for ONLINE-OFFLINE replication
- ☹ **Additional database installation needed** for application requiring write access (split of OFFLINE database)

Context:
The need to set up another DB instance to support the Tier1 replication environment is a special case for CERN

Summary

- Evaluated different options for database replication to Tier1s.
- **Streams11g** remains the most suitable technology at present
  - **Operational** concerns outweighed by performance benefits and **familiarity** with the technology.
- However, **Active DataGuard** is extremely interesting
  - As part of the overall **export** process
  - To improve **redundancy** in the **online** environment and when disaster recovery site is implemented

ACFS 11.2 tests



Oracle RAC Oracle Database
Application
Application
Oracle ASM Files for Oracle Database
Oracle ASM Cluster File System (ACFS)
Third Party File System
Oracle ASM Dynamic Volume Manager (ADVM)
Oracle Automatic Storage Management
Operating System

Introduction

- Automatic Storage Management (ASM)
  - Oracle's **cluster file system and volume manager** for Oracle databases
- ASM Dynamic Volume Manager (ADVM)
  - new feature in Oracle Clusterware 11.2
  - **volumes** are implemented **as ASM files**
  - exposed to OS as block devices
- ASM-based Cluster File System (ACFS)
  - new in Oracle 11.2
  - built **on top of ADVM** volumes
  - can be used **cluster-wide** or **single-node** only

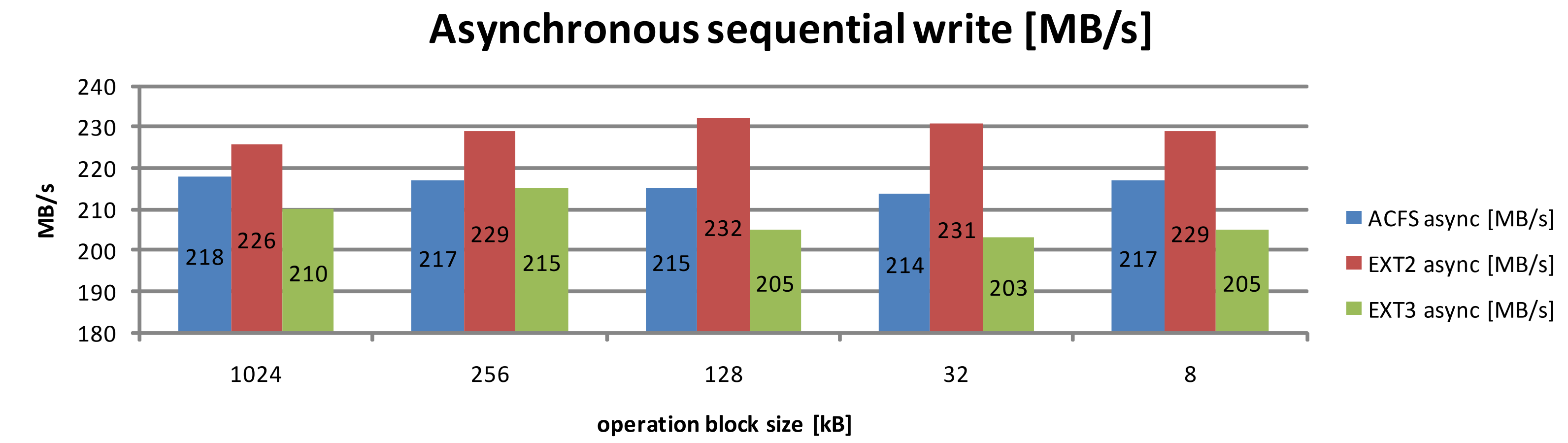
ACFS use cases at CERN

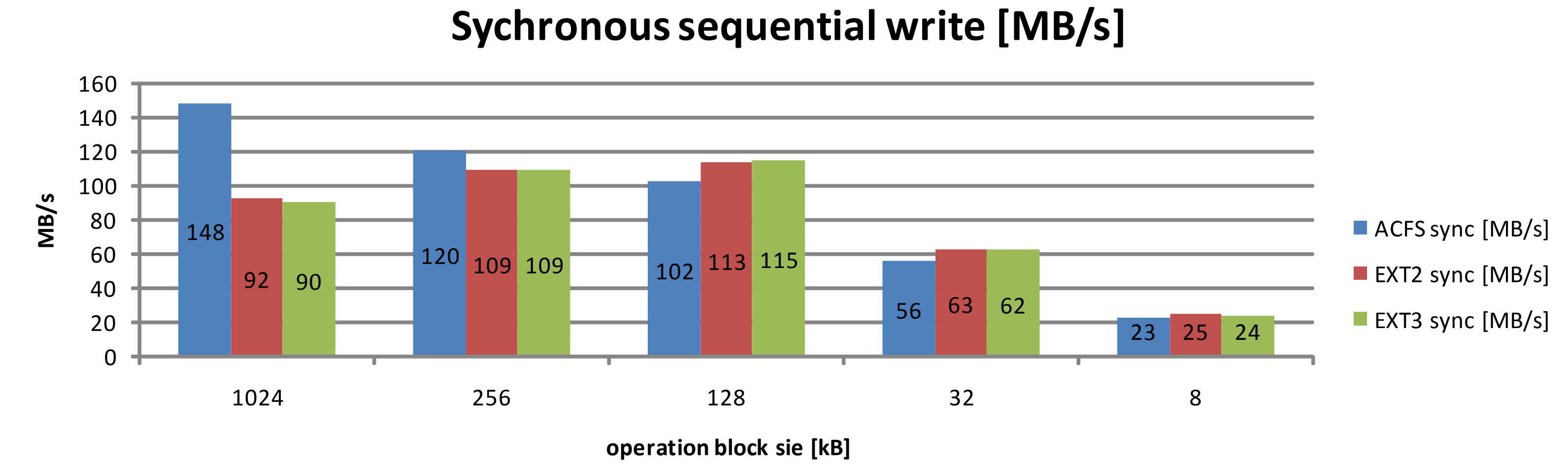
- ACFS is used in **production** at CERN
  - General purpose cluster file system for backup & monitoring cluster – **fast and reliable**
  - **Repository** of oracle binaries
  - **Temporary storage** for large exports/imports
- Potential usages
  - Automatic Diagnostic Repository (ADR)
  - Export/import directory for each cluster DB

Tests conducted

- Tests description
  - Comparing ACFS, ext2, ext3 and encrypted ACFS (AES 192-bit)
  - ADVM used in all tests
- Compared operations
  - Sequential write (synchronous and asynchronous)
  - Sequential read (synchronous)
  - File system block write, rewrite and read; file creation and deletion speed
  - Multithread tests

Write test results in our environment





ext2 shown as reference for raw performance but not usable in large scale environments

Summary

- ACFS usage at CERN
  - Positive experience
    - Currently used to provide cluster file system for our custom DB monitoring
  - Positive results from performance tests
    - More tests in progress

Outreach

CERN openlab

Presentations

- ***"ACFS under scrutiny"*** Luca Canali and Dawid Wojcik, UKOUG, Birmingham

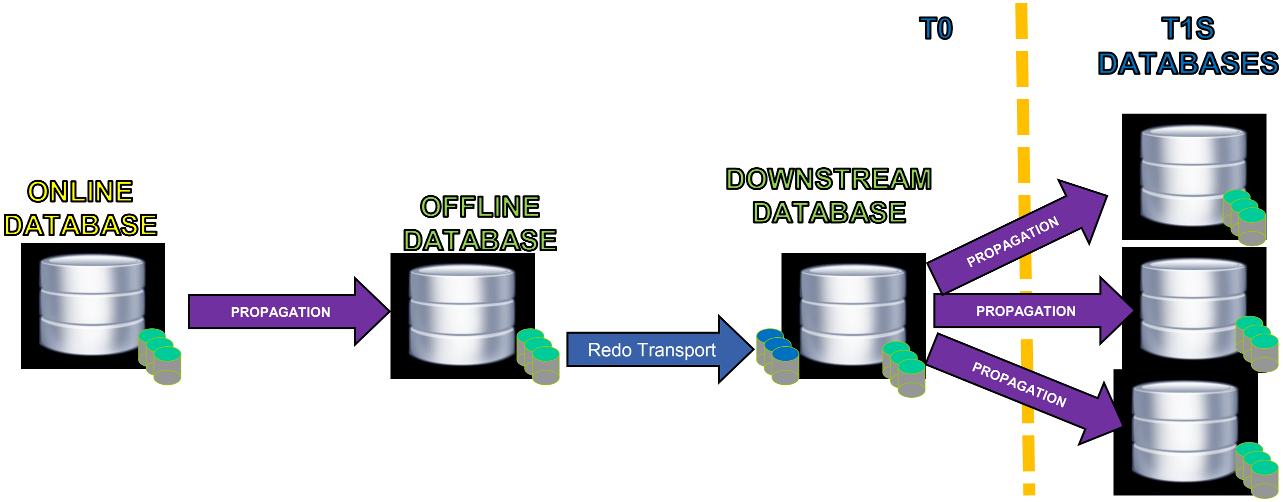
- ***"Data Lifecycle Management Challenges and Techniques, a user's experience***" Luca Canali and Jacek Wojcieszuk, UKOUG, Birmingham
- ***"Distributed Database Workshop"*** @CERN in November. Presentation from Michael Smith (Oracle)

- **In preparation:**
  - Press realease
  - Reference call
  - Presentation in iCSC

Questions?

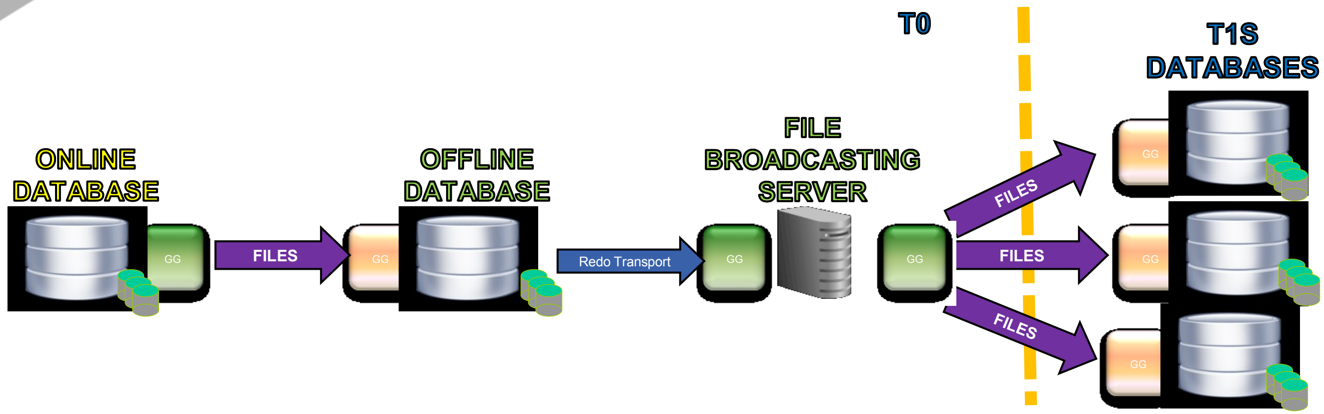
Backup slides

Streams11gR2 replication at all Tiers



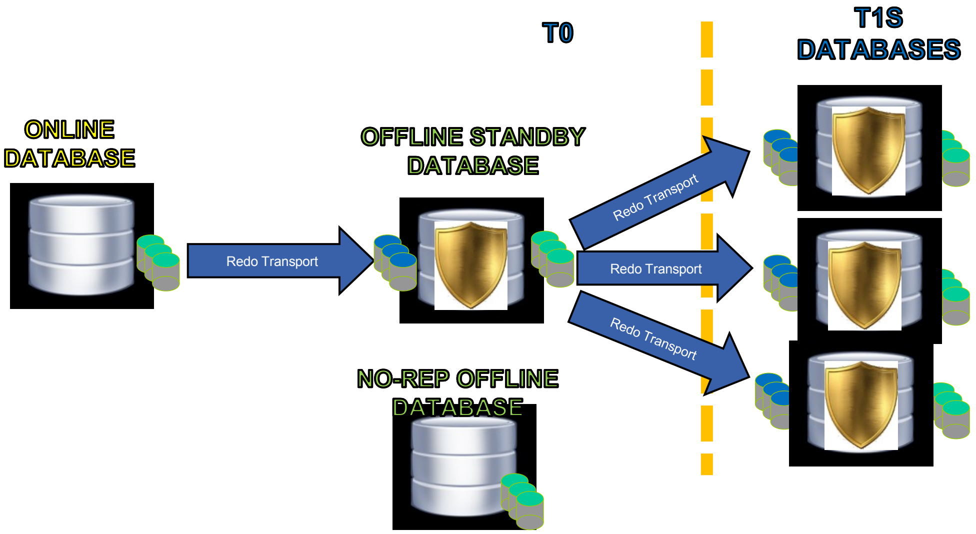
- ☺ Best practices identified – a **lot of experience**
- ☺ **Good monitoring** for distributed streams deployment (strmmon, EM)
- ☹ **Additional hardware** required (downstream capture) to isolate the source database
- ☹ **Recovery** of replica **requires coordination** between T1s and T0

GoldenGate11g replication at all Tiers



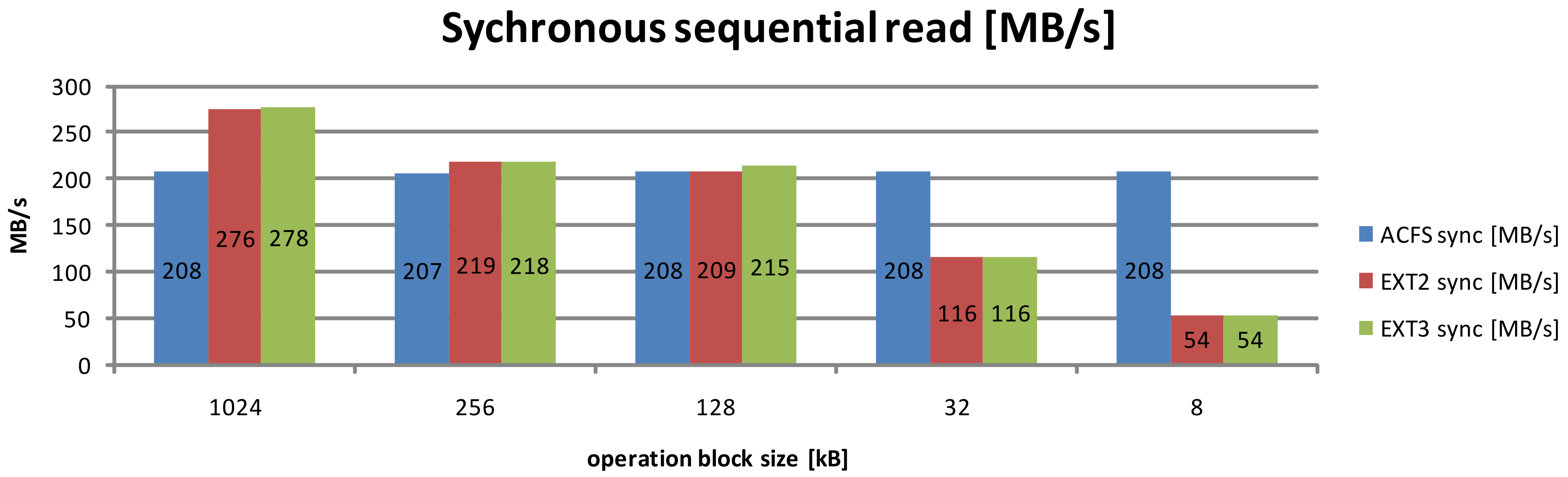
- ☺ Easier **maintenance**
  - ☺ **No** side **effects** on source when **target** is **down** - no split of replication required
  - ☺ Trail files can be used for T1 recovery – **no coordination needed** from T0
- ☹ **Short** in-house **experience**
- ☹ **No monitoring** for distributed environment available
- ☹ **No performance improvement** for our replication environments in comparison with Streams

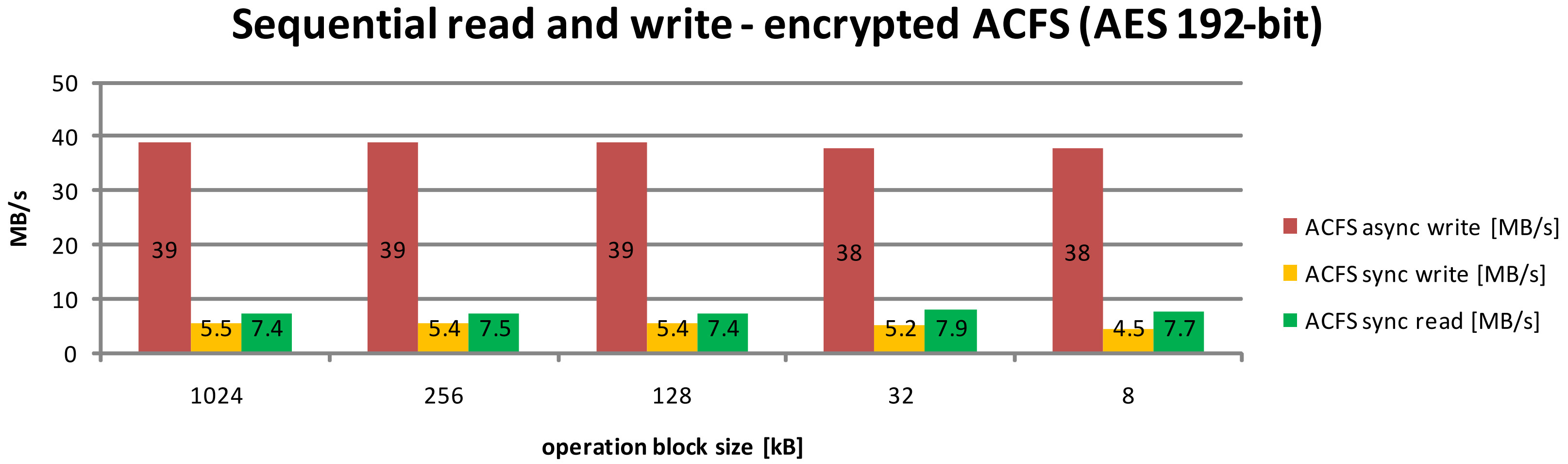
Active DataGuard 11gR2



- ☺ **Lower maintenance effort** (physical replication)
- ☺ **Less impact of users activity** on the replication
- ☹ **Same version of DB required** at all Tiers
  - Coordination of interventions becomes critical
- ☹ **Additional database installations needed** for no replicated data (split of OFFLINE)

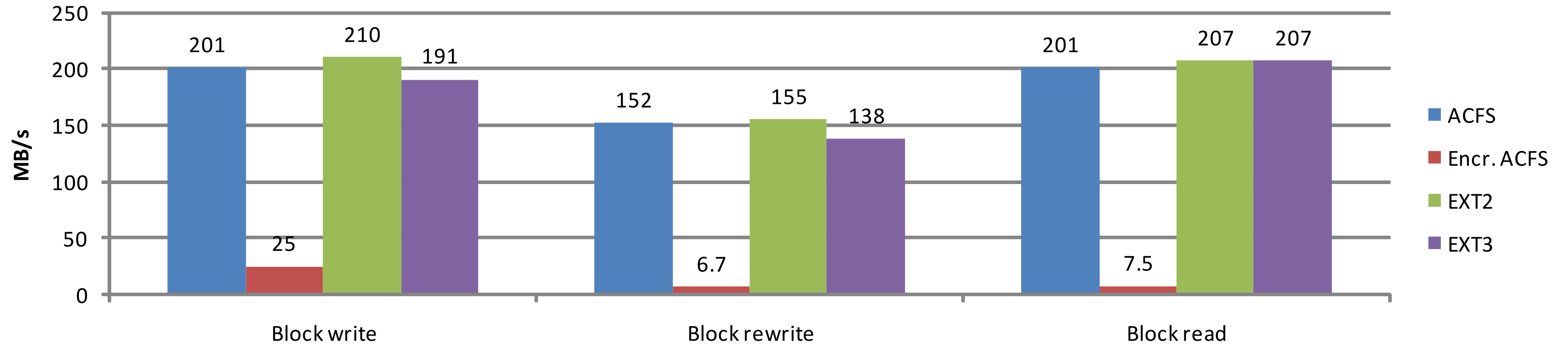
Read and write results in our enviroment



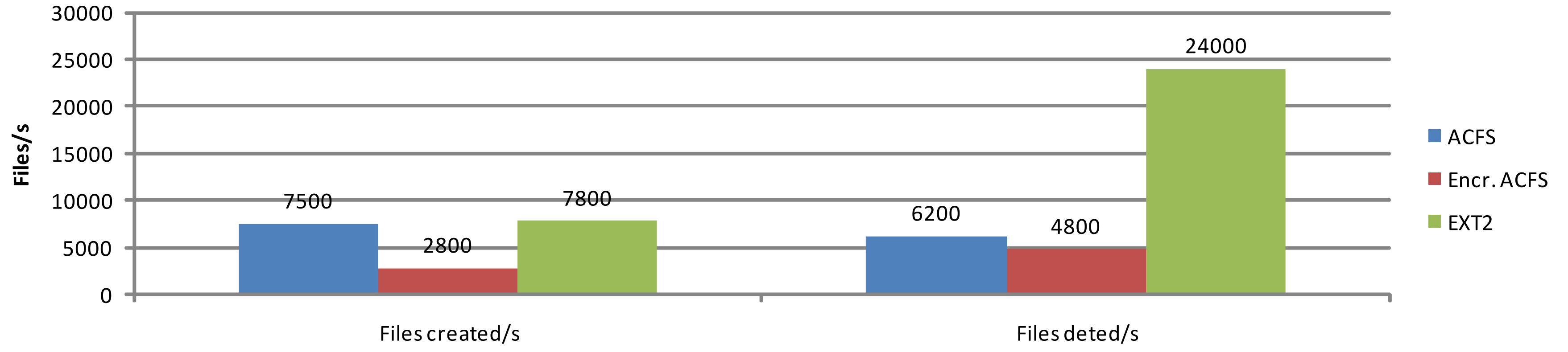


bonnie++ test results in our environment

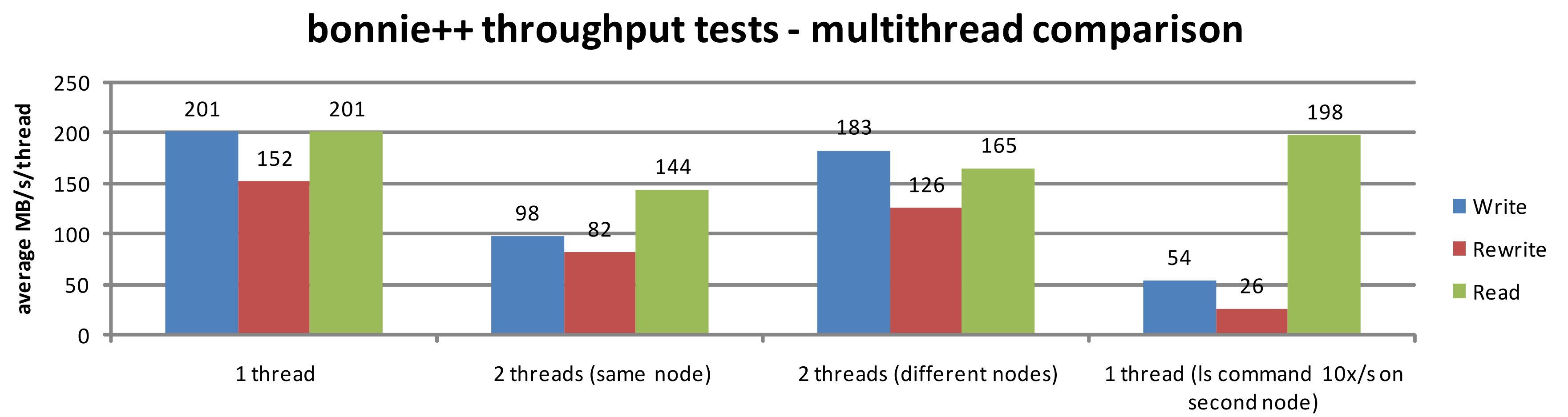
bonnie++ throughput tests

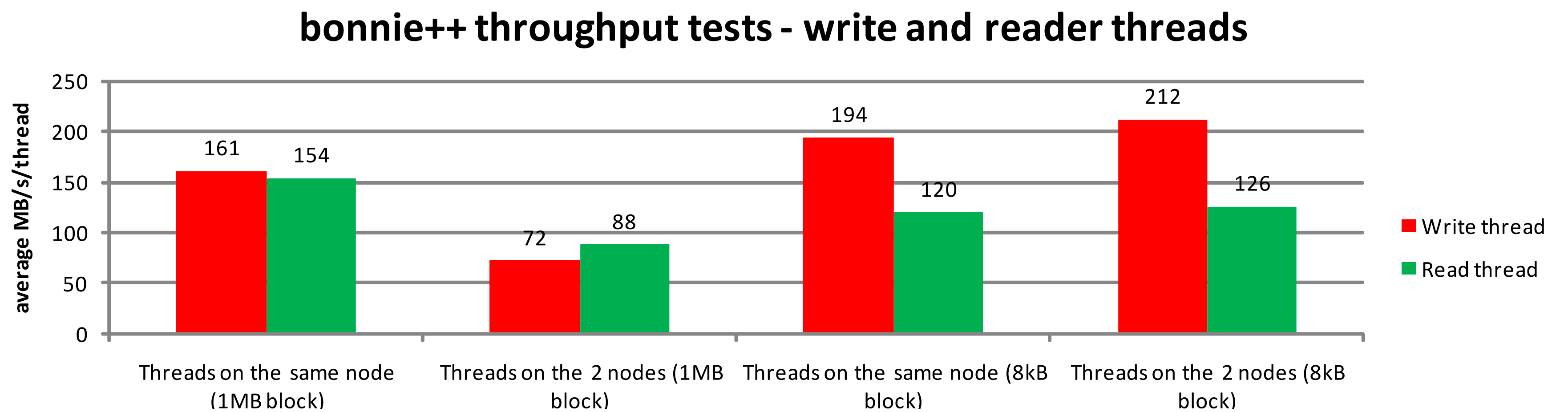


bonnie++ random file creation test



Multithread test results in our environment





Tests results

- Asynchronious sequential write [Block Size=1kB]
  - Ext2 226MB/s, **ACFS** 218 MB/s, Ext3 210 MB/s
- Synchronous sequential write
  - **ACFS** 148 MB/s, Ext2 92 MB/s, Ext3 90 MB/s
- Synchronous sequential read
  - Ext3 278 MB/s, Ext2 276 MB/s, **ACFS** 208 MB/s
- ACFS sequentail I/O encrypted
  - async writes 39 MB/s, sync writes 5.5 MB/s, reads 7.4 MB/s
- Random file creation speed
  - Ext2: 7800 files /s, **ACFS** 7500 files/s, Encr. ACFS 2800 files /s
- File deletion speed
  - Ext2: 24000 files /s, **ACFS** 6200 files /s, Encr. ACFS 4800 files /s
- Multithread test
  - 2 threads running on the same node
    - Write: 196 MB/s
    - Read: 288 MB/s
  - 2 threads running on a different nodes
    - Write 366 MB/s
    - Read 330 MB/s