Enabling High-Performance Database Applications on Supercomputing Architectures

M. Modani, M.A. Johnston, D. Moss, K. Jordan

Natural Disasters ...

Oasis

IBM Research

- **Oasis is framework that allows insurance companies estimate the loss they could incur due to natural disasters**
- Developed as free open-source architecture by Oasis LMF ltd.
- Oasis specifies format input data should take and how the calculations are performed – ***transparency***
- The aim is to build an open community that can innovate on different aspects of loss-modelling e.g. calculations, software, hardware etc.
- The vision is for an open market place of data for the calculations
- Oasis is supported by a consortium of interested parties including Llyods of London who commissioned this short-term project

Oasis Website: http://www.oasislmf.org/

How Oasis Works: Architecture

IBM

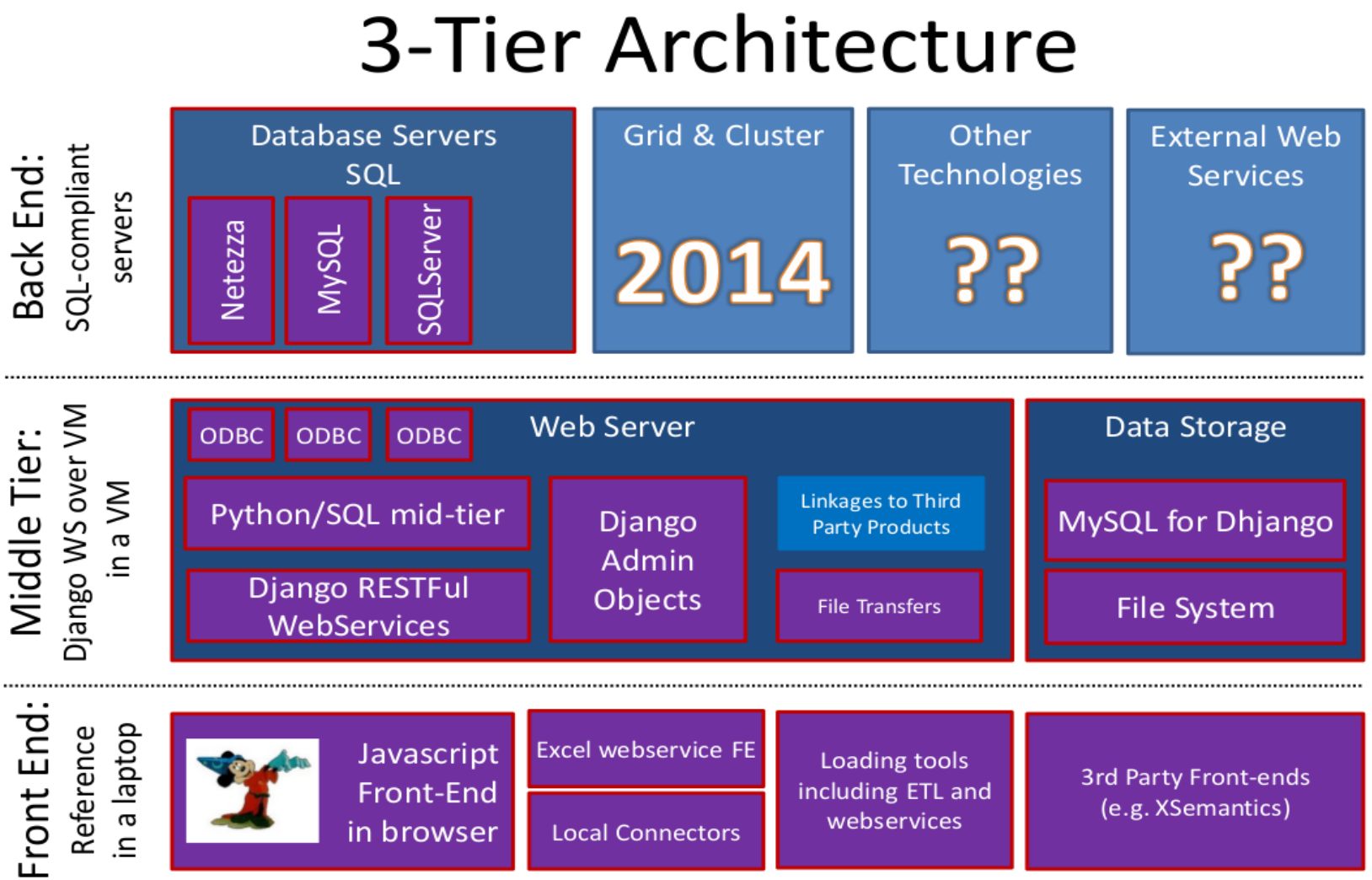


Image Credit: Oasis LMF

How Oasis Works: Kernel Calculation

IBM Research

- The core computation of Oasis (Back-end Tier) consists of two steps
- Damage CDF
  - Generate a cumulative probability distribution function (CDF) for each combination of catastrophic event, area and building type
  - Example: flood, uk postcode, 3 bed 1950 semi
  - Each of these combinations is called a **FACT**
  - The number of FACTs is determined by the number of policies the insurer has (its *exposures*)
- Ground-Up Loss (GUL)
  - Uniform Monte-Carlo sampling of the CDF
  - Aim: Calculate the expected loss for the FACT
  - The potential loss can be dominated by events at tail of CDF

How Oasis Works: Kernel Calculation

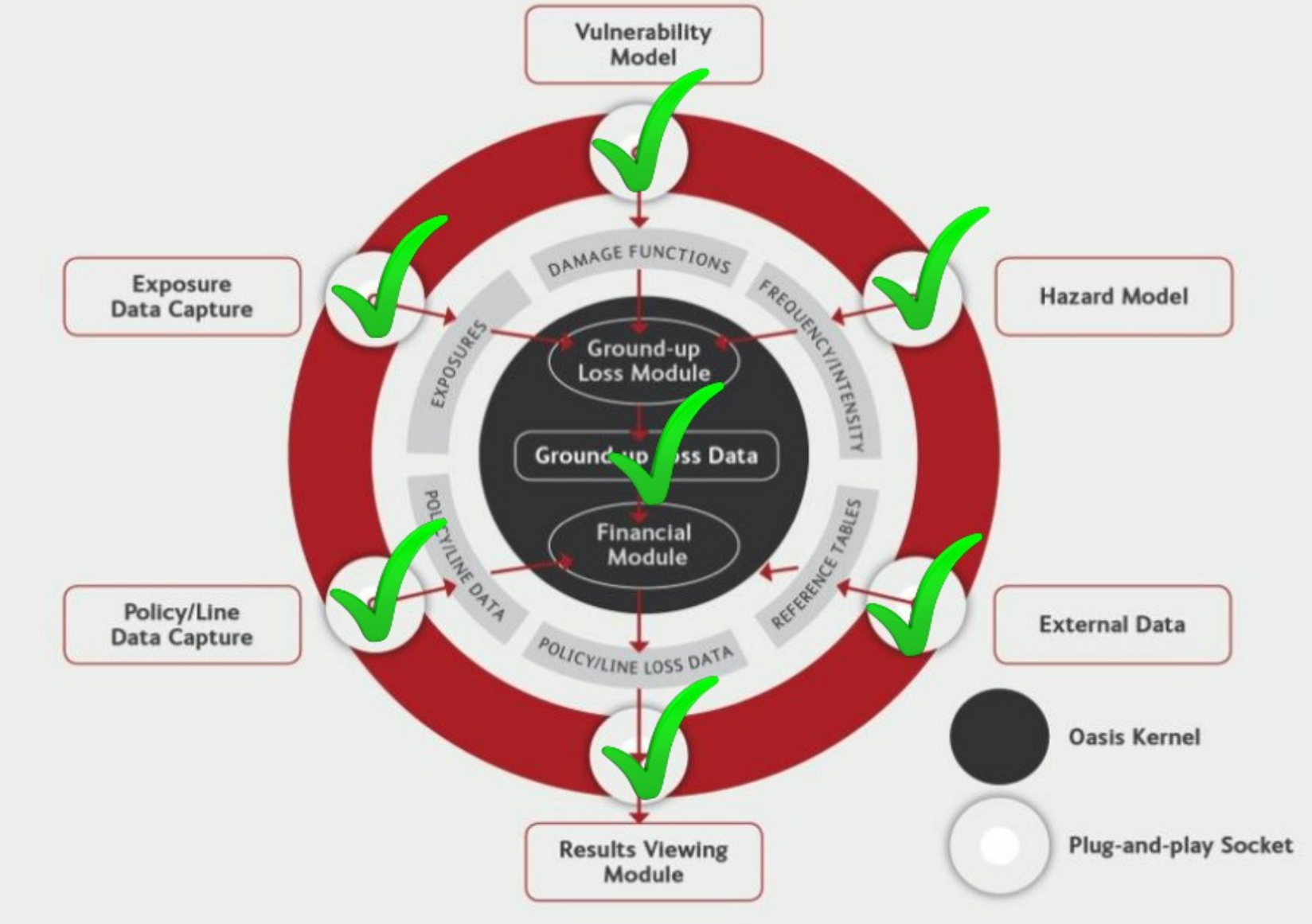


Image Credit: Oasis LMF

Why HPC?

IBM Research

- Damage Step
  - UK Flood Model Data-Set: 77.1 million FACTs
  - 100 points per CDF = 7.71 billion table rows
    - At 4ms per CDFgeneration = 3.5 days to calculate just CDFs
- GUL Step
  - For each CDF you sample N times to calculate loss –
    - bigger N → better estimation → longer time
- Both steps tax Compute and I/O
- However all FACTS are independent – good for parallelisation
  - Process groups of FACTs, called ***chunks***, simultaneously.

IBM Research

Project Aim: **Understand how the Oasis LMF could exploit HPC**

Hartree Centre
Science & Technology Facilities Council

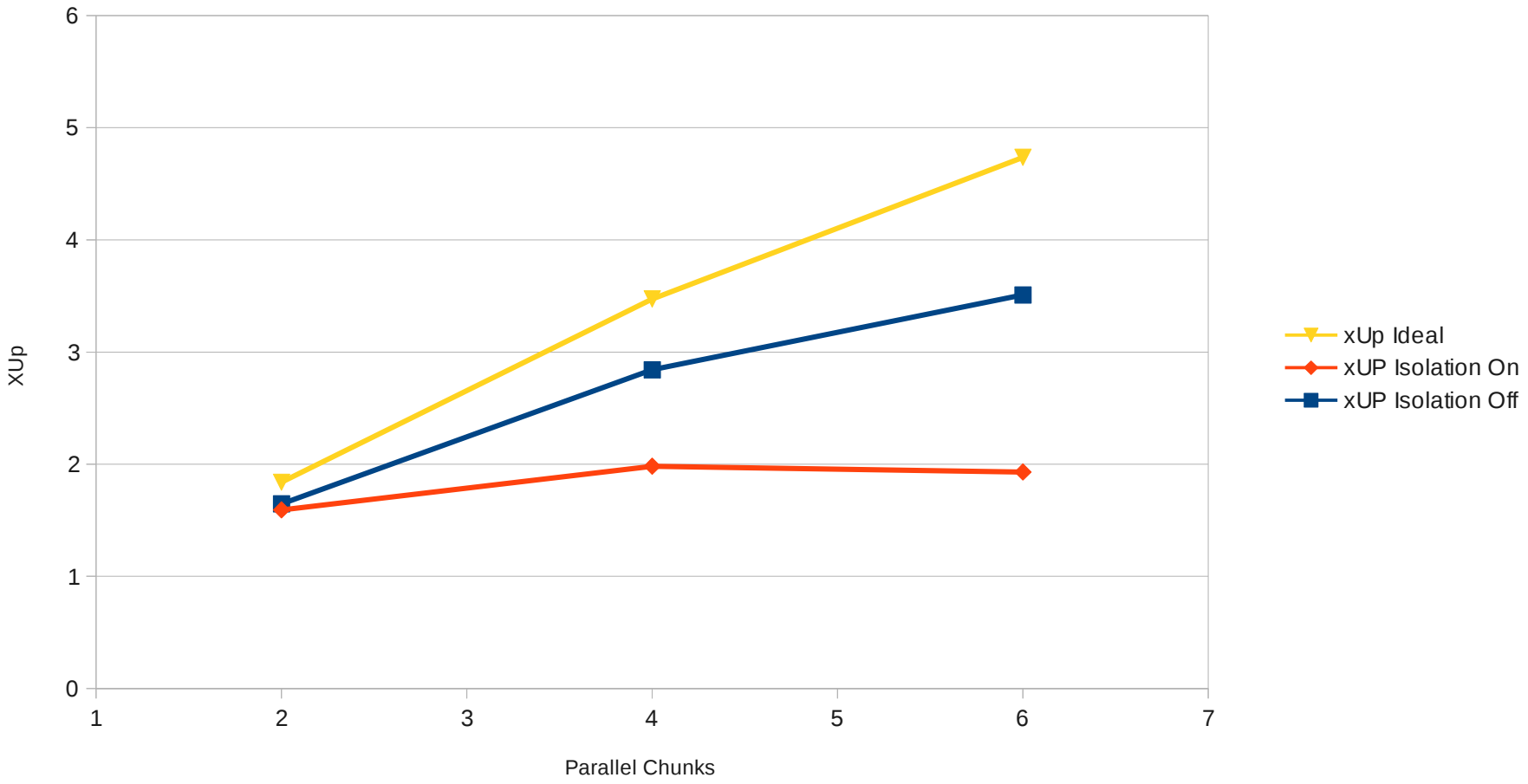
“Harnessing the power and unlocking the potential of high performance computing”

Challenges

IBM Research

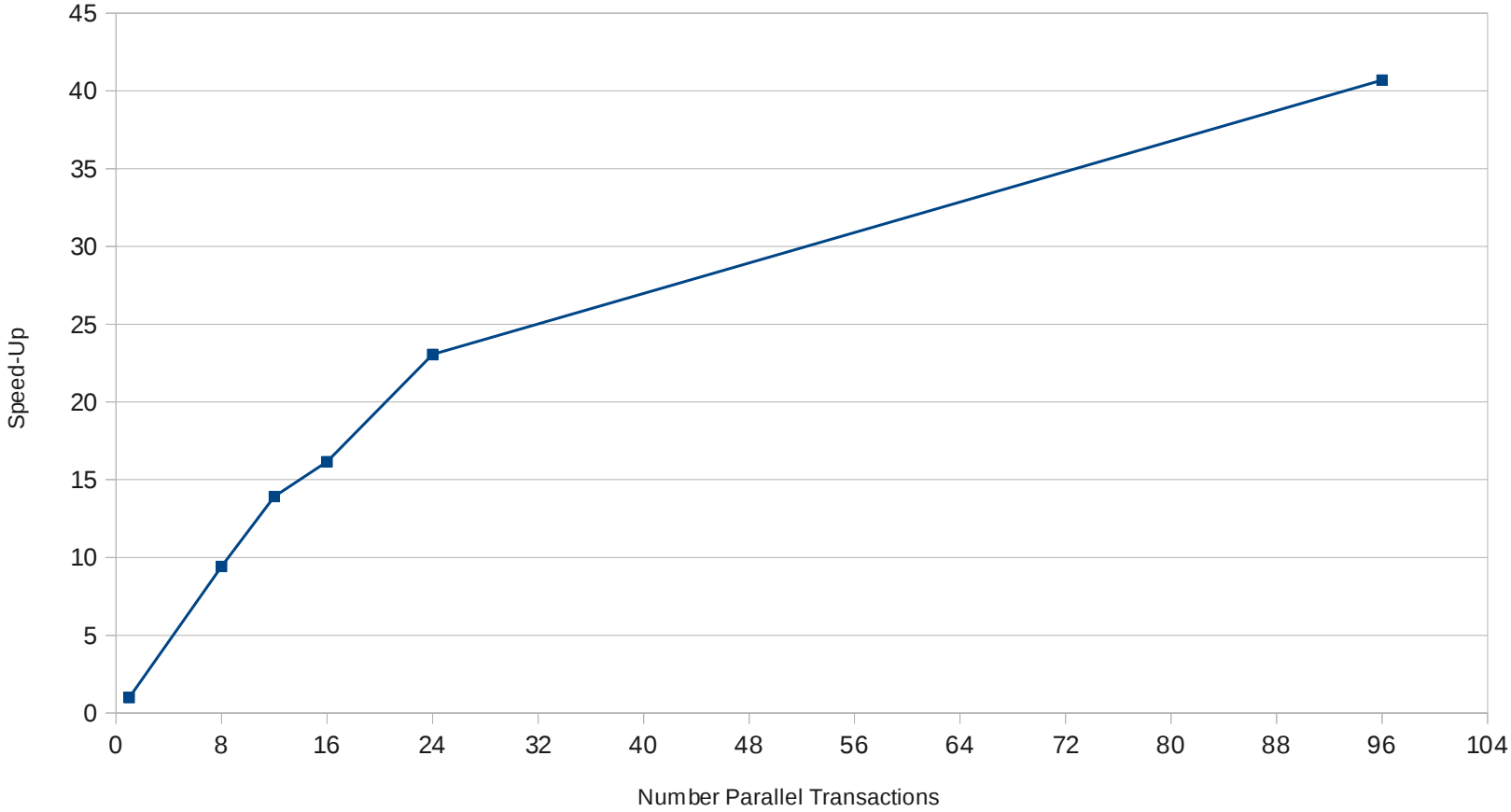
- Oasis Kernel is written in SQL ...
- Oasis architecture built around dedicated database server
  - Assumes there is an always on database server which you can connect to
- Supercomputers are shared and resource managed
  - Can't run dedicated database server on cluster nodes
  - Can't access cluster nodes external to system to run from mid-tier
- Oasis developed for SQLServer, Netezza, and MySQL (third choice and least optimized)
  - MySQL only option for deployment on iDataplex
  - No explicit parallelism implemented
- Computational scientists and not database experts!

Results – Single Node Transaction Parallelism



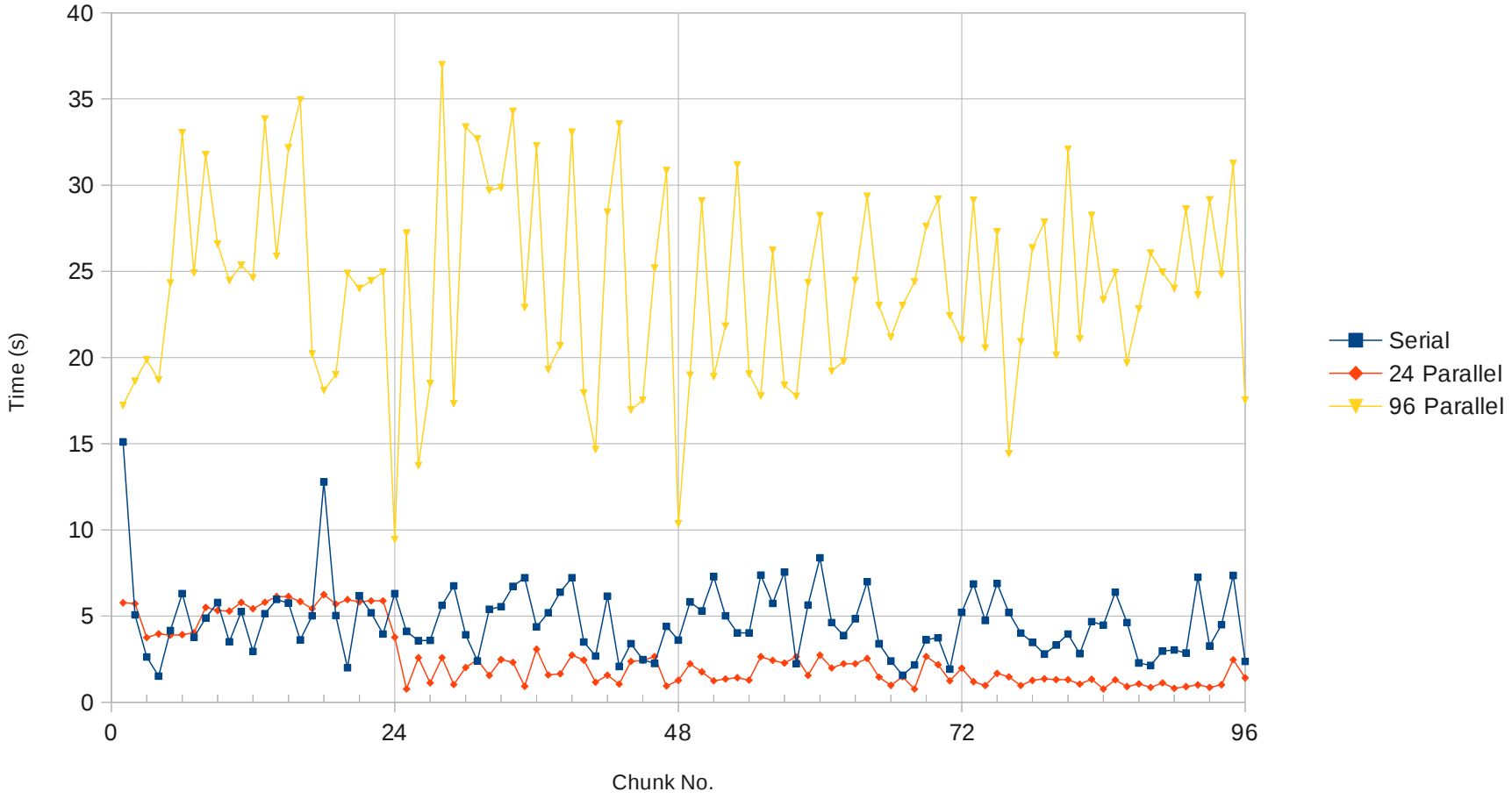
Speed-Up obtained by processing chunks in parallel on a single node for the Damage step. The possible xUP is dependent on the transaction isolation level.

Results – Multi Node Transaction Parallelism



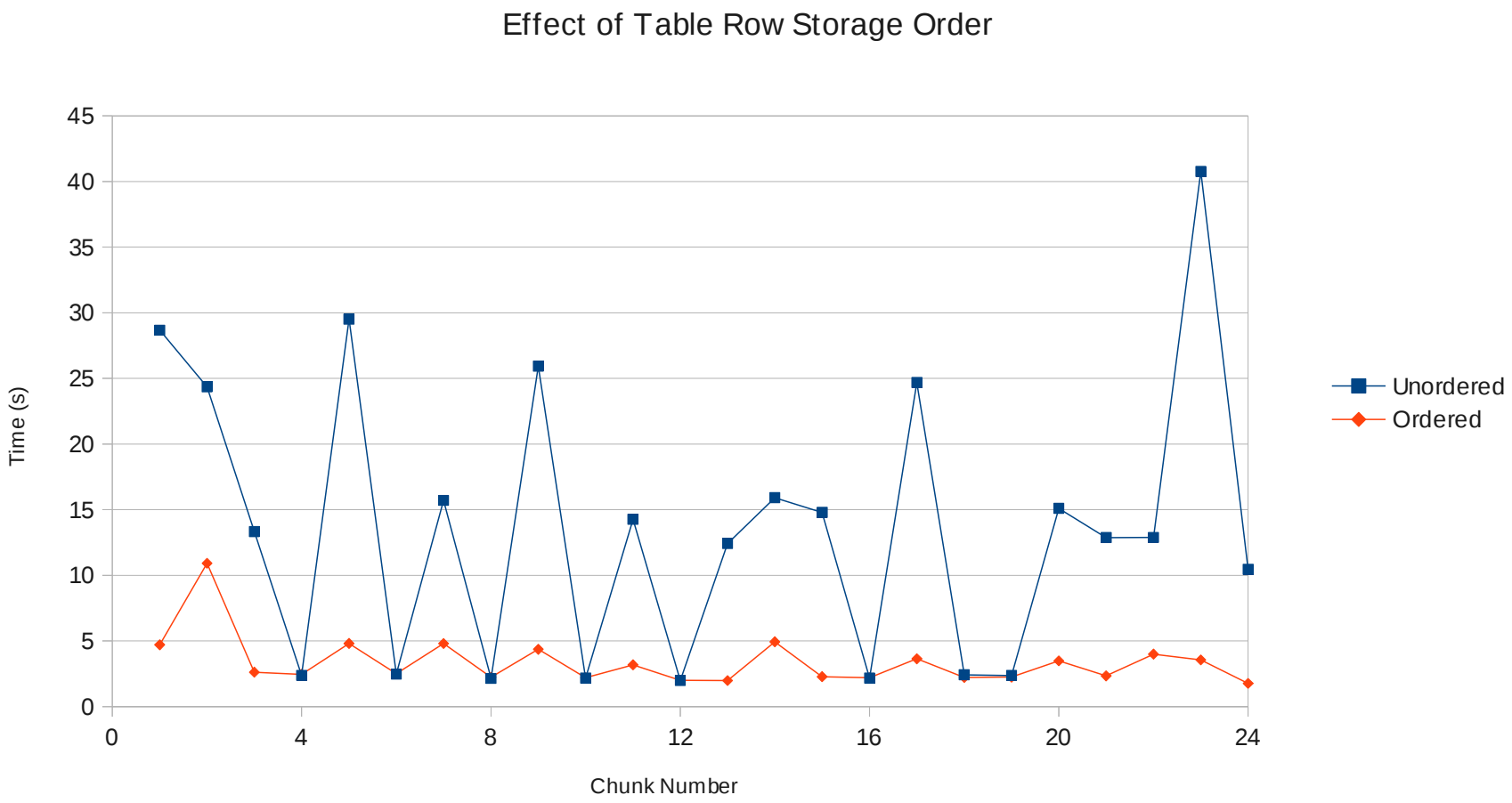
Parallelisation of the Damage step using a 24 SQL node MySQL Cluster. Up to 24 parallel transactions only distributed parallelism used. For 96 parallel transactions a hybrid scheme was used (4 transactions in parallel per cluster node) – A max xUP of 40.7x was obtained.

Results – Barriers To Scaling



Lock contention due to parallel CREATE/DROP statements limits scaling. The lines are the time to execute all CREATE/DROP statements in each transaction executed with different numbers in parallel. At 96 parallel transactions this time increase significantly

Results – I/O Characteristics



The order that SQL table rows are stored on disk is important for harnessing GPFS capabilities.

Results – Summary

IBM Research

- Deployed Oasis on Blue Wonder iDataplex using MySQL and MySQL Cluster
  - First deployment of Oasis on MySQL Cluster and first MySQL Cluster on shared iDataplex
- Analysed and optimised MySQL query performance
  - e.g. CDF creation from 11ms to 1ms
- Analysed and optimised impact of GFPS/IO
- Parallelised Oasis Calculation across multiple iDataplex nodes
  - 49x speed up over base MySQL GUL time and 41x CDF time
    - Calculating CDF for 0.5 million FACTS from 35 minutes to 52 seconds
  - GUL time comparable to Netezza (33 microseconds to 42 microseconds per CDF).
- Identified key barriers to scaling Oasis Kernel further on HPC architectures.

Conclusions

IBM Research

- MySQL is an effective execution engine for Oasis on HPC nodes
- MySQL Cluster provides an effective method of harnessing distributed memory parallelism but scaling bottlenecks limit achievable performance
- Launching a database server on shared HPC cluster tricky
  - Requires specific interfaces to be built to be used from Oasis Mid-Tier
- Likely a non-SQL implementation of Oasis Kernel could provide much higher performance than possible with SQL
  - The ACID properties provided by relational databases have a negative impact on performance