Contribution ID: **378** Type: **Talk**

AI-Assisted Optimization of Large-Scale Climate Data Transfers in South African Research Infrastructure

Wednesday, 3 December 2025 12:00 (20 minutes)

AI-Assisted Optimization of Large-Scale Climate Data Transfers in South African Research Infrastructure

CHPC Conference 2025, Cape Town

Abstract

Background and Motivation

The transfer of large-scale scientific datasets between South African research facilities represents a critical bottleneck in computational research workflows. Climate modeling datasets of the Global Change Instituite, Wits University , as odf Aug2025, are just over 540TB over 3 users, particularly from the Conformal-Cubic Atmospheric Model (CCAM).Optimized transfer strategies between the Centre for High Performance Computing (CHPC) and the Data Intensive Research Initiative of South Africa (DIRISA) storage systems are thus necessary for resilient data flows between HPC, storage and local analsysis compute facilities.Current data transfer tools such as Globus Connect identify bottlenecks within data flow circuits, however manual command line iRODS interfaces present significant challenges for reliable data transfer through AI-assisted optimization."

Methodology: AI-Assisted Development

This work presents a systematic application of artificial intelligence tools (Claude Code) to develop filesystem-aware transfer optimization solutions. The AI-assisted development process generated three complementary tools in under 4 hours of development time:

1. **Performance benchmarking script** for systematic testing of 24-core Data Transfer Node configurations
2. **Resilient transfer wrapper** with exponential backoff retry logic and comprehensive verification
3. **Lustre-aware optimization engine** that dynamically analyzes filesystem striping patterns and adjusts transfer parameters

Technical Innovation: Lustre Filesystem Integration

The core innovation lies in automated Lustre striping analysis using lfs getstripe commands, coupled with dynamic parameter optimization. The system automatically detects:
- Stripe counts and sizes for optimal thread allocation
- Object Storage Target (OST) distributions for concurrency planning
- File size patterns for buffer optimization
- Directory structures for efficient batch processing

Performance Results

Test Dataset: 189TB CCAM climate modeling installation (ccam_install_20240215)
- **Source**: CHPC Lustre filesystem (/home/jpadavatan/lustre/)
- **Destination**: DIRISA iRODS storage (/dirisa.ac.za/home/jonathan.padavatan@wits.ac.za/)

Validation Testing (67MB, 590 files):
- **Success Rate**: 100.0% (590/590 files transferred successfully)
- **Transfer Performance**: 0.95 GB/hour sustained throughput
- **Reliability**: Zero failed transfers with comprehensive verification
- **Peak Performance**: 10.41 MB/s maximum transfer rate
- **Optimization**: Automatic 8-thread, 64MB buffer configuration

Scalability Analysis:
- **Small datasets** (41-67MB): 100% success rate, 4-11 MB/s
- **Medium datasets** (17GB): Structure-preserving transfers completed
- **Large datasets** (20-34TB per directory): Systematic optimization applied

AI Development Impact

The AI-assisted approach delivered significant advantages:
- **Development Speed**: Complete toolchain developed in <4 hours vs. estimated weeks for traditional development
- **Code Quality**: Production-ready tools with comprehensive error handling and logging
- **Documentation**: Auto-generated usage examples and architectural documentation
- **Iterative Improvement**: Real-time debugging and enhancement based on performance feedback

South African Research Infrastructure Impact

Immediate Benefits:
- Enables systematic transfer of 189TB CCAM climate datasets to long-term DIRISA storage
- Provides reusable toolchain for other large-scale data transfers in SA research community
- Demonstrates AI-assisted development methodology for research computing infrastructure

Broader Applications:
- Astronomical data transfers (MeerKAT, SKA precursor datasets)
- Genomics datasets from National Health Laboratory Service
- Earth observation data from SANSA and international collaborations
- General HPC-to-archive workflows across CHPC user community

Technical Contributions

1. **First documented AI-generated, striping-aware transfer optimization** for African research infrastructure
2. **Open-source toolchain** available at: https://github.com/padavatan/chpc-irods-transfer-tools
3. **Comprehensive auditing framework** with source validation, performance tracking, and efficiency analysis
4. **Systematic methodology** for AI-assisted research infrastructure development

Future Work and Scalability

Planned extensions include:
- Integration with CHPC job scheduling systems for automated large-scale transfers
- AI-assisted optimization for emerging storage technologies
- Performance modeling for petabyte-scale climate datasets

Conclusion

Initially, the twin challenges of troubleshooting the Globus bottelneck and creating manaul workflow setup of irods scripts presented as techncical complexities requiring considerable investment in troubleshooting

effort. This work demonstrates that AI-assisted development can dramatically accelerate research infrastructure optimization while maintaining production-grade reliability. The 100% success rate achieved in validation testing, combined with comprehensive filesystem-aware optimization, provides a foundation for systematic large-scale data management in South African research computing.

The methodology offers promise to the local HPC research community and other institutions facing similar data transfer challenges and represents a paradigm shift toward AI-augmented research infrastructure development.

Technical Implementation Stack

AI Development Platform:
- **Claude Code** (claude.ai/code): Primary AI development assistant for code generation, debugging, and optimization
- **Development Time**: <4 hours vs. estimated 2-3 weeks traditional approach (10-15x productivity improvement)

Core Technologies:
- **Languages**: Python 3.9.6, Bash scripting
- **HPC Infrastructure**: CHPC 24-core DTN systems, Lustre parallel filesystem, DIRISA iRODS storage
- **Specialized Tools**: iRODS iCommands (iput, ils), Lustre client tools (lfs getstripe), GNU Parallel
- **Version Control**: Git, GitHub CLI, collaborative development workflow

Performance Analysis Framework:
- **Custom benchmarking**: Transfer rate analysis with variance tracking
- **Comprehensive auditing**: Source/destination validation, file integrity verification
- **Real-time monitoring**: Speed measurements, efficiency metrics, optimization recommendations

AI-Human Collaboration Model:
- **AI Contributions**: 2,597+ lines of production code, comprehensive error handling, filesystem-aware algorithms, automated documentation
- **Human Contributions**: Domain expertise, performance validation, requirements specification, production integration
- **Result**: Production-grade reliability with rapid development cycles

This technical stack demonstrates the practical implementation of AI-assisted research infrastructure development, providing a replicable methodology for other HPC environments.

Keywords: Artificial Intelligence, High-Performance Computing, Data Transfer Optimization, Lustre Filesystem, iRODS, South African Research Infrastructure, Climate Modeling, CHPC, DIRISA

Authors: Jonathan Padavatan[1], Mthetho Sovara[2], Claude (AI Assistant)[3]
[1] University of the Witwatersrand, Global Change Institute
[2] CHPC
[3] Anthropic AI

Contact: jonathan.padavatan@wits.ac.za
Repository: https://github.com/padavatan/chpc-irods-transfer-tools

Presenting Author

Jonathan Padavatan

Email

jonathan.padavatan@wits.ac.za

Student or Postdoc?

No. Not a student nor Postdoc.

Institute

University of the Witwatersrand

Registered for the conference?

Yes

CHPC User

Yes

CHPC Research Programme

ERTH1200

Primary authors: Mr PADAVATAN, Jonathan (University of the Wiitwatersrand); Mr SOVARO, Mthetho (CHPC)

Presenter: Mr PADAVATAN, Jonathan (University of the Wiitwatersrand)

Session Classification: DIRISA

Track Classification: DIRISA