Integrating End-to-End Exascale SDN into the LHC Data Distribution Cyberinfrastructure

Jonathan Guiang
Aashay Arora
Diego Davila
John Graham
Dima Mishin
Thomas Hutton
Igor Sfiligoi
Frank Würthwein
jguiang@ucsd.edu
aaarora@ucsd.edu
didavila@ucsd.edu
jjgraham@ucsd.edu
dmishin@ucsd.edu
hutton@sdsc.edu
isfiligoi@sdsc.edu
fkw@ucsd.edu
UC San Diego
La Jolla, California, USA

Tom Lehman
Xi Yang
Chin Guok
tlehman@es.net
xiyang@es.net
chin@es.net
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Harvey Newman
Justas Balcas
newman@hep.caltech.edu
jbalcas@caltech.edu
Caltech
Pasadena, California, USA

ABSTRACT

The Compact Muon Solenoid (CMS) experiment at the CERN Large Hadron Collider (LHC) distributes its data by leveraging a diverse array of National Research and Education Networks (NRENs), which CMS is forced to treat as an opaque resource. Consequently, CMS sees highly variable performance that already poses a challenge for operators coordinating the movement of petabytes around the globe. This kind of unpredictability, however, threatens CMS with a logistical nightmare as it barrels towards the High Luminosity LHC (HL-LHC) era in 2030, which is expected to produce roughly 0.5 exabytes of data per year. This paper explores one potential solution to this issue: software-defined networking (SDN). In particular, the prototypical interoperation of SENSE, an SDN product developed by the Energy Sciences Network, with Rucio, the data management software used by the LHC, is outlined. In addition, this paper presents the current progress in bringing these technologies together.

CCS CONCEPTS

• **Networks → Network management**; *Network algorithms*; **Network services**; **Network performance evaluation**; • **Information systems → Data management systems**; • **Applied computing** → *Operations research.*

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3535134

KEYWORDS

exascale, data distribution, software defined networking

ACM Reference Format:
Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi, Frank Würthwein, Tom Lehman, Xi Yang, Chin Guok, Harvey Newman, and Justas Balcas. 2022. Integrating End-to-End Exascale SDN into the LHC Data Distribution Cyberinfrastructure. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491418.3535134

1 INTRODUCTION

Like much of the scientific community, the LHC at CERN [4] is barrelling towards an exascale era. Projections (Fig. 1) indicate that, after the High Luminosity upgrade, the CMS experiment alone will produce roughly 0.5 exabytes of data per year. This data will be organized, as it is now, into datasets consisting of hundreds to millions of O(GB) files. Moreover, it is generally accessed by two different parties:

- **Users:** thousands of physicists and students around the world need to retrieve and process CMS data, driving millions of small data transfers
- **Production:** CMS produces real and simulated detector data and distributes it around the world, driving hundreds of large data transfers

Crucially, this implies that CMS data movement is entirely dominated by the production data transfers, which are centrally managed. Furthermore, given the global nature of both user and production data transfers, National Research and Education Networks

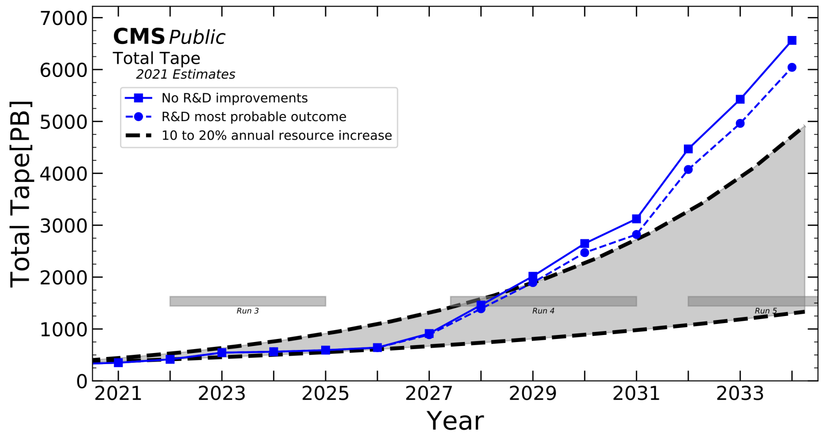

Figure 1: CMS tape usage estimates from 2021 made for the CMS contribution to the LHC Committee (LHCC) November review of HL-LHC computing and common software [5]. Two scenarios are considered: a baseline (solid blue) that does not include any improvement and the most probable outcome (dashed blue) of the ongoing R&D activities.

(NRENs) become critical cyberinfrastructure for CMS data distribution. However, every NREN is opaque to CMS, as they are to every other domain science they serve, resulting in highly variable performance–unpredictably oscillating between excellent and poor. This kind of variable network performance already poses a challenge for operators coordinating petascale production transfers, but a looming exascale era–meaning more and possibly larger production transfers–makes a solution all the more imperative [1, 8]. If CMS were able to instead guarantee a large fraction of available bandwidth for high-priority production transfers, they could be achieved on a predictable timescale. Meanwhile, users could share just a small portion of the remaining bandwidth without much performance loss. This exactly could be achieved by software-defined networking. Moreover, CMS is currently unable to easily distinguish site issues from network issues, so when data movement is slowed, the underlying problem cannot be quickly addressed. If, however, CMS had an account of what was promised from the network, that account could be compared to the site diagnostics it already maintains in order to immediately identify the culprit. In other words, SDN bandwidth allocations can be thought of as exact "promises" made by the network that it is then held accountable for maintaining. Therefore, a realized SDN product would provide predictable and accountable network services, both of which CMS currently lacks and could greatly benefit from in the HL-LHC era. This paper presents the prototypical interoperation of such an SDN product, SENSE, with the data-movement management system, Rucio, used by CMS.

CMS is one of two general-purpose physics experiments at the LHC, the other being ATLAS. Notably, ATLAS also projects exascale HL-LHC data production [1, 8], yet this paper is written entirely in the context of CMS. The work is nevertheless relevant to both collaborations, since ATLAS also uses Rucio–albeit a separate instance from the one used by CMS. The data-movement workflow and cyberinfrastructure employed by ATLAS is also fundamentally similar to that used by CMS, and thus it faces many of the same challenges that this work aims to address.

1.1 Rucio: Scientific Data Management

Rucio [3] is an open-source software framework that is currently used to manage all of the data produced by CMS and ATLAS, which primarily involves orchestrating production data transfers. This is done by defining "rules," where each rule represents the replication of one or more datasets from one site to another. They are each assigned a priority, then put on a stack–sorted by that priority–to be transferred by FTS [2]. Importantly, Rucio can be configured to allow these priorities to be changed mid-transfer.

1.2 SENSE: SDN for End-to-End Networked Science at the Exascale

The Energy Sciences Network (ESNet), the NREN leveraged by CMS and ATLAS in the United States, is developing an SDN product called SENSE [7], with the titular aim of providing end-to-end SDN for exascale scientific data movement. In particular, SENSE builds on existing SDN work by providing enhanced interactivity through an intent-based interface. This allows for management of the network at the same level as how instruments, compute, and storage are managed. Specifically, it lets users intuitively customize Layer 2 and Layer 3 services to build guaranteed-bandwidth "links" between two sites.

The site configuration (Fig. 2) that enables the construction of SENSE links is a critical contribution of this work, so the basic operating principle is described here in more detail. Every site must provide multiple gobally routable IPv6 subnets to serve as potential SENSE link endpoints. Each of these IPv6 subnets is associated with a "redirector" that can direct traffic to one of the data servers connected to it, where each server has equal access to the site's filesystem. The exact role and performance of these components is defined by the XRootD architecture, which is used by CMS and ATLAS [6]. In any case, a site configured in this way can then support multiple SENSE links between itself and another site. This therefore enables the fundamental action of the Rucio-SENSE interoperation prototype: establishing guaranteed network service for a given data

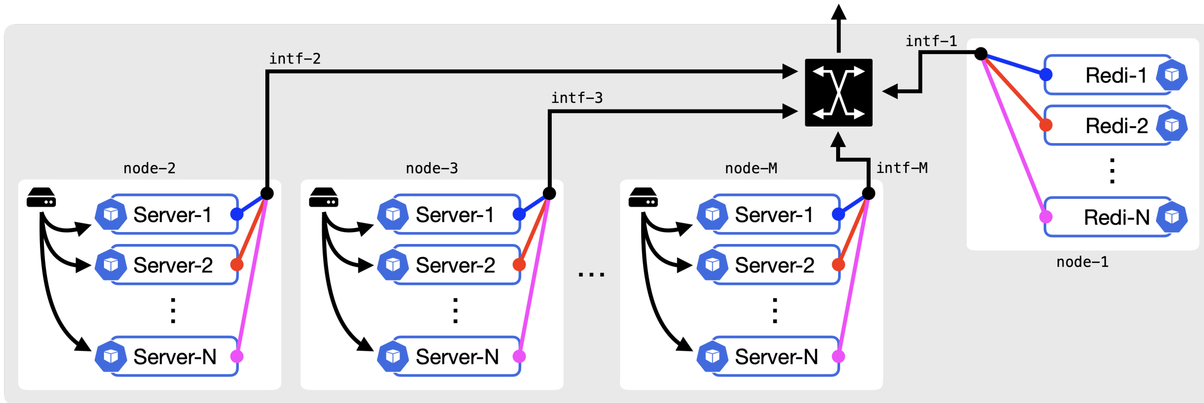

Figure 2: A generic SENSE site configuration. A "redirector" (e.g. *Redi-1*) listens to a specific IPv6 address, from a unique IPv6 subnet, and directs incoming traffic to one of the data servers connected to it (e.g. *Server-1*). Each data server has equal access to the site's filesystem.

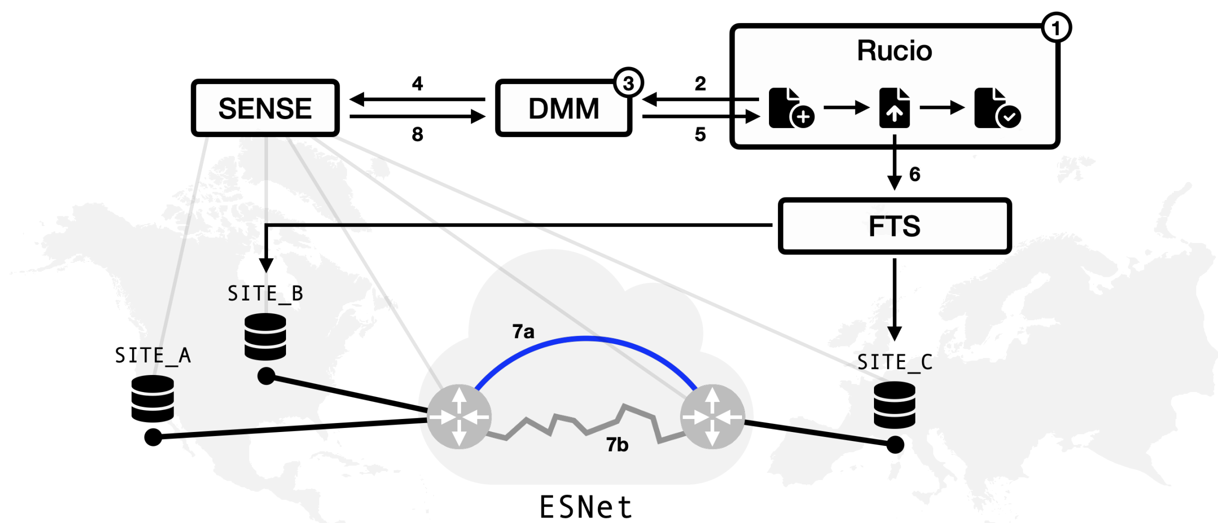

Figure 3: Simplified diagram of the Rucio-SENSE interoperation workflow, with numbered steps: (1) a rule is initialized; (2) Rucio sends transfer description to DMM; (3) DMM translates Rucio request into SENSE provision; (4) DMM sends provision request to SENSE; (5) DMM sends a source IPv6 and destination IPv6 to Rucio; (6) Rucio injects IPv6s from previous step into the FTS request; (7) Either (a) SENSE builds dedicated service or (b) default to best effort service; (8) SENSE sends service metadata to DMM.

transfer between two sites. In addition, this configuration scheme defines an important, configurable constraint; that is, the maximum number of links that SENSE can construct between any two sites is limited by whichever site has fewer configured IPv6 subnets.

2 RUCIO-SENSE INTEROPERATION MODEL

The central goal of this work is to produce a Rucio-SENSE interoperation model that enables Rucio operators to prioritize certain transfers, then see those priorities reflected in the allocation of network resources. Moreover, this should have a minimal impact on the current implementation and operation of Rucio. To this end, a Data Movement Manager (DMM) is introduced to perform the crucial task of translating Rucio requests into SENSE provisions and returning the results, keeping a bulk of the functionality required for the incorporation of SENSE capabilities separate from the Rucio codebase. Thus, DMM serves as a keystone for the Rucio-SENSE interoperation model explored in this initial work (Fig. 3):

(1) A Rucio operator initializes a rule with some priority which requests one or more dataset transfers, where each transfer may involve a different pair (source and destination) of sites
(2) Rucio sends the following data to DMM for each transfer:
  - Total transfer size
  - Source site
  - Destination site

  - Priority
(3) DMM processes the data from Rucio:
  (a) If the transfer has no priority, immediately place it on best effort service (skip steps below)
  (b) Reserve an IPv6 address at the source and destination site
  (c) Compute the bandwidth provision (i.e. promise) appropriate for the transfer priority
(4) DMM requests a new promise from SENSE that implements the provisioning from (3c), reprovisioning existing promises where appropriate
(5) DMM sends the IPv6 addresses it reserved to Rucio
(6) Rucio injects the IPv6 addresses into the FTS request
(7) SENSE takes one of the following actions:
  (a) Begin the construction of a new guaranteed-bandwidth link
  (b) Do nothing; the transfer will be provided best effort service
(8) SENSE sends identifying metadata for the link back to DMM

Several of these steps offer the opportunity for further optimization. In particular, it is clear that a future implementation may see the integration of DMM into Rucio such that the operations in steps (2) through (5) can be implemented to better handle a large number of transfers. Before then, however, DMM provides an isolated testbed in which the fundamental design of the interoperation model can be prototyped. Step (3c) is of particular interest, because it implements the bandwidth provisioning–the central deliverable of this work. The provisioning decision could be designed, for example, to allow Rucio operators to schedule transfers: e.g. *move Dataset A to Site B in one week*. Alternatively, the decision could pack the maximum available bandwidth between two sites as a function of each transfer's priority. In any case, this interoperation model provides a flexible testbed for evaluating these ideas and producing concrete metrics on their scalability, practicality, and efficiency.

3 CURRENT STATUS

At the time of writing, the software side of this work is nearly ready to test. The basic functionality required for the interoperation model is already available in SENSE, and the first version of DMM has been written. In addition, Kubernetes-based XRootD configurations for the source and destination sites have been prepared; this work requires a particularly unusual XRootD configuration, but functionality has been rigorously tested and confirmed. The hardware side of this work is also nearing completion, despite the typical pandemic-related technology supply chain delays. Testbeds with specific networking capabilities are being assembled at Caltech and UCSD. Meanwhile, SENSE-specific site configurations are being piloted at Caltech and are nearly ready to bring into the current version of SENSE. Once everything is assembled, basic tests will begin immediately to show that all of the moving parts work in concert. Finally, functionality explored in this prototype is expected to be gradually moved into use during biennial LHC data challenges, with the next starting in 2023, and from there into production.

As most global collaborations in physics and astronomy share the same software stack as ATLAS and CMS, much of the same network infrastructure, and share many of the same facilities, especially in Europe and Asia, it is expected that the work described here will eventually be adopted widely across global science collaborations, once proven successful in production by its initial adopters.

ACKNOWLEDGMENTS

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-2030508, OAC-1841530, OAC-1836650, MPS-1148698, and PHY-1624356. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DE-SC0015528, DE-SC0016585, and FP-00002494. Finally, this work would not be possible without the significant contributions of collaborators at ESNet, Caltech, and SDSC.

REFERENCES

[1] Johannes Albrecht et al. 2019. A Roadmap for HEP Software and Computing R&D for the 2020s. *Computing and Software for Big Science* 3, 1 (20 Mar 2019), 7. https://doi.org/10.1007/s41781-018-0018-8
[2] A A Ayllon, M Salichos, M K Simon, and O Keeble. 2014. FTS3: New Data Movement Service For WLCG. *Journal of Physics: Conference Series* 513, 3 (Jun 2014), 032081. https://doi.org/10.1088/1742-6596/513/3/032081
[3] Martin Barisits, Thomas Beermann, Frank Berghaus, et al. 2019. Rucio: Scientific Data Management. *Computing and Software for Big Science* 3, 1 (09 Aug 2019), 11. https://doi.org/10.1007/s41781-019-0026-3
[4] CERN. 2022. CERN Homepage. https://cern.ch Accessed: 2022-04-29.
[5] CMS Collaboration. 2022. CMS Offline and Computing Public Results. https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults Accessed: 2022-04-29.
[6] Alvise Dorigo, Peter Elmer, Fabrizio Furano, and Andrew Hanushevsky. 2005. XROOTD-A highly scalable architecture for data access. https://xrootd.slac.stanford.edu/presentations/xpaper3_cut_journal.pdf Accessed: 2022-04-01.
[7] Inder Monga, Chin Guok, John MacAuley, Alex Sim, Harvey Newman, Justas Balcas, Phil DeMar, Linda Winkler, Tom Lehman, and Xi Yang. 2020. Software-Defined Network for End-to-end Networked Science at the Exascale. *Future Generation Computer Systems* 110 (2020), 181–201. https://doi.org/10.1016/j.future.2020.04.018
[8] J. Zurawski, D. Brown, B. Carder, E. Colby, E. Dart, K. Miller, et al. 2021. *2020 High Energy Physics Network Requirements Review Final Report.* Technical Report LBNL-2001398. Lawrence Berkeley National Laboratory. https://escholarship.org/uc/item/78j3c9v4