

Integrated Research Environment Briefing

Sunshine Week

March 16, 2017

John Fattaleh
Center for Economic Studies

Motivation

- To support a high-level of research needed for the Census Bureau's mission, we need a state of the art computing infrastructure.
- The existing system was not scalable and not flexible enough to meet researcher needs.
- The solution is the Integrated Research Environment (IRE).

Potential Benefits

- Streamline the process to leverage expertise in the Federal Statistical Research Data Centers (FSRDC).
- Project infrastructure provides a direct link between DMS and researcher access.
- Potential for the Census Bureau to save storage space (by reducing duplication of static data).
- Software tools championed by one group for IRE would potentially be available to all researchers, in particular the use of new open source software.

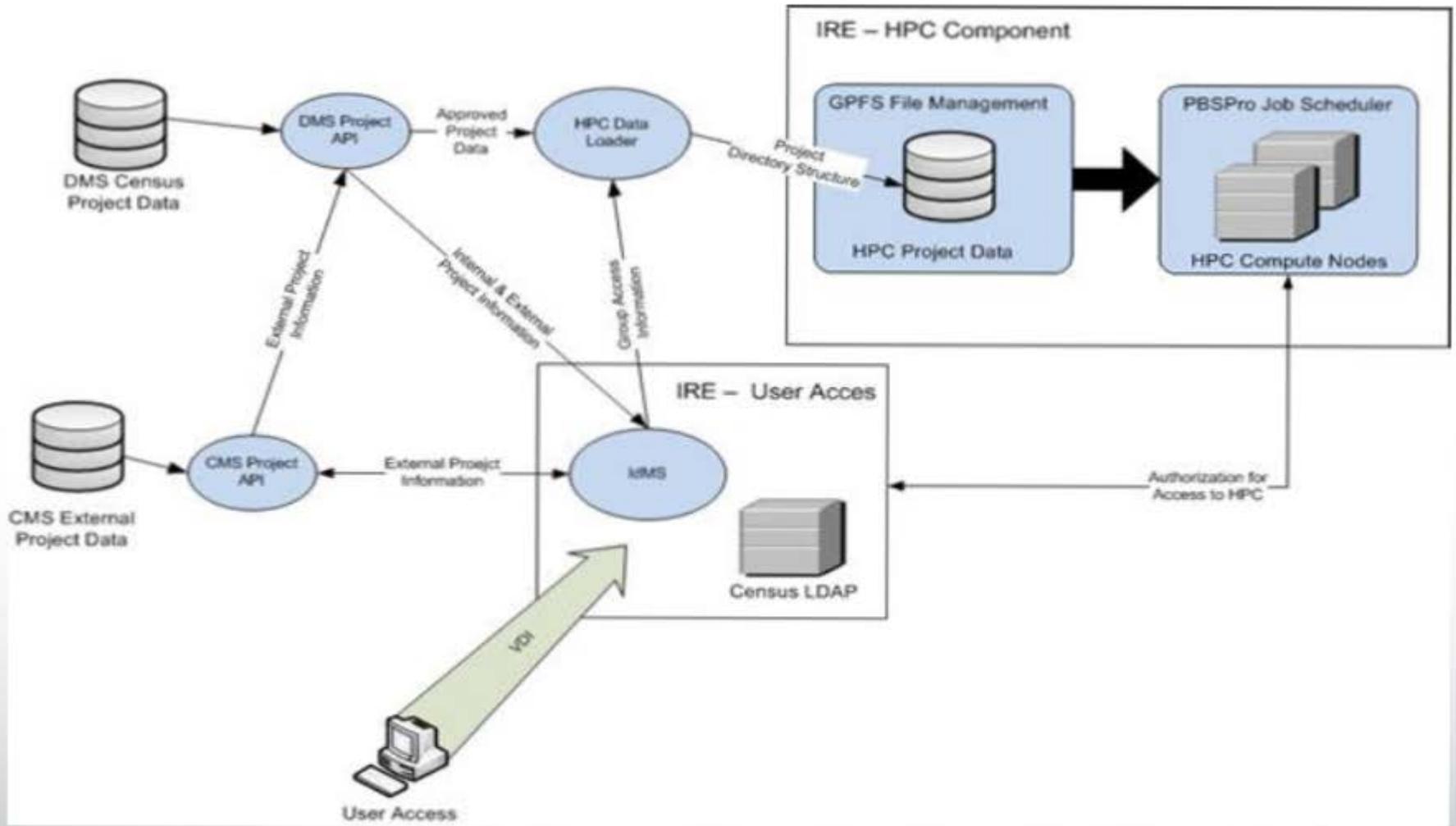
Joint Effort – R+M and IT

- 8 Divisions spread across three directorates...
- ADRM
 - Center for Economic Studies (CES)
 - Center for Statistical Research and Methodology (CSRM)
- ADITCIO
 - Information Systems Support and Review Office (ISSRO)
 - Computer Services Division (CSvD)
 - Office of Information Security (OIS)
 - Telecommunication Office (TCO)
 - LAN Technology Support Office (LTSO)
- Application Development and Services Division (ADSD)

Timeline

- **Fall 2011:** ADRM White Paper on Research Computing
- **November 15, 2012:** Project Charter Signed
- **September 29, 2014:** Complete Prototype Evaluation Project and Established Production Architecture
- **December 15, 2016:** Authority to Operate
- **FY17:** Migration and expansion

Architectural Diagram



IRE Components

- “High Performance” Computing
 - General Parallel File System (GPFS)
 - Can scale up to large amounts of CPU, RAM, and SAN
 - Can have heterogeneous OS and server sizes
- Virtual Desktop Infrastructure
 - Access to internal network
 - Enterprise supported
- REST API
 - CES Management System (CMS, external projects)
 - Data Management System (DMS, all projects)
 - Identity Management System (IdMS, all projects)

High Performance Computing

- Enterprise Resources (corporately funded by Census)
 - 3 small compute nodes (24x256)
 - 1 large compute node (40x512)
- Program Resources (funded by specific programs)
 - 6 new small compute nodes
 - 17 small compute nodes (will migrate)
 - 1 large compute node (will migrate)
 - 10 large compute nodes (32x768, planned)

Demand for IRE

- Storage:
 - Static Data Warehouse ~75TB
 - Project space ~.5 Petabytes (PB)
- Users:
 - Federal Statistical Research Data Centers – 800 researchers
 - Census Bureau
 - National Center for Health Statistics
 - Bureau of Labor Statistics
 - Other agencies planning to join...
 - Census Bureau Staff – 300 researchers

Software Tools Available

- NX Server/NX client (communication software)
- Job Scheduler - Portable Batch System (PBS) Professional (PBSPro)
 - SAS
 - Stata
 - R
 - Matlab
 - Tomlab
 - KNITRO
 - Anaconda Python
 - Many others...

Comparable Environment

- **FDA Computing Cluster**
 - 350+ servers
 - 3500+ CPUs
- **Three Largest Centers**
 - Jointly funded effort
 - Shared governance over resources

Migration

- **Phase 1:** Research Data Center cluster external projects
- **Phase 2:** Research 1 cluster internal projects
- **Phase 3:** Research 2 cluster internal projects

Future enhancements?

- Cloud Computing
- HADOOP
- InfiniBand
- Graphics Processing Units (GPUs)

Questions/Feedback

- John.A.Fattaleh@census.gov
- Adley.Kloth@census.gov
- Shawn.D.Klimek@census.gov