Paper SD11

Maximizing Longevity and Saving Costs Through Refactoring of Statistical Computing Environments

Vaithi Bharath, Bayer, Whippany, US
Srinath Viswanathan, Bayer, Whippany, US
Srinivas Veeragoni, Bayer, Whippany, US

Abstract

At Bayer, we have a home-grown Statistical Computing Environment (SCE), built using C & Java on Linux, that has served our needs for the past 20 years. Maintenance for the SCE has become more complex due to outdated technology, retiring engineers, and a shortage of relevant skilled professionals in the market. While evaluating the market for SCE offerings, we have also evaluated the option to refactor the current platform to optimize code, introduce automation, generate LLM-powered documentation, and migrate the platform to the cloud with file systems optimized for supporting advanced analytics while keeping preserving the business functionality and improving the user experience. In this paper we will walk you through our refactoring journey and how this could result in cost savings and better operational efficiency in the near term while we build the enterprise strategy for the longer term with the integration of SAS Viya and Posit components.

Introduction

This paper details our journey through the refactoring process, highlighting the potential significant cost savings and operational efficiencies. Additionally, it outlines our strategic vision for the future, which includes the integration of SAS Viya and Posit components, to build a more resilient and scalable enterprise strategy. As we navigate this transformation, we aim to set a new standard for statistical computing environments that align with contemporary technological advancements and business needs.

Problem Statement

As Bayer's home-grown SCE, it has become increasingly difficult to maintain due to outdated technology, the retirement of experienced engineers, and a shortage of relevant skilled professionals in the market. This legacy system, which has served our needs for the past 20 years, now faces significant challenges in terms of operational efficiency and scalability.

The primary problem is to determine the most effective approach to modernize the SCE, ensuring it meets current and future business needs. This involves evaluating whether to adopt a new market offering or to refactor the existing platform. Key considerations include:

Code Optimization: Enhancing the current codebase to improve performance and maintainability.

Automation: Introducing automated processes to reduce manual intervention and errors.

Documentation: Generating comprehensive documentation using large language models (LLMs) to facilitate knowledge transfer and system understanding.

Cloud Migration: Transitioning the platform to a cloud-based architecture, with optimized file systems to support advanced analytics.

The solution must also ensure the preservation of existing business functionalities while achieving significant cost savings and better operational efficiency in the short term. Additionally, it should align with Bayer's long-term enterprise strategy, including the integration of SAS Viya and Posit components, to create a more resilient and scalable SCE.

Transforming Bayer's Statistical Computing Environment

In response to the challenges posed by our aging SCE, Bayer embarked on a transformative journey to modernize its legacy systems, SWAN and Macumba. These applications, built on a 15-year-old platform, required a comprehensive overhaul to meet contemporary technological standards and business needs. Our primary objectives were to optimize performance, enhance user experience, and establish a scalable infrastructure that supports future innovations.

Migration and Refactoring
Over the course of 8-10 months, we successfully migrated SWAN and Macumba to Amazon Web Services (AWS), leveraging Infrastructure-as-Code to streamline environment provisioning. This transition not only modernized our infrastructure but also laid the groundwork for cloud-native platform architecture, along with continuous integration and continuous deployment (CI/CD) capabilities, essential for the upcoming integration with SAS Viya.

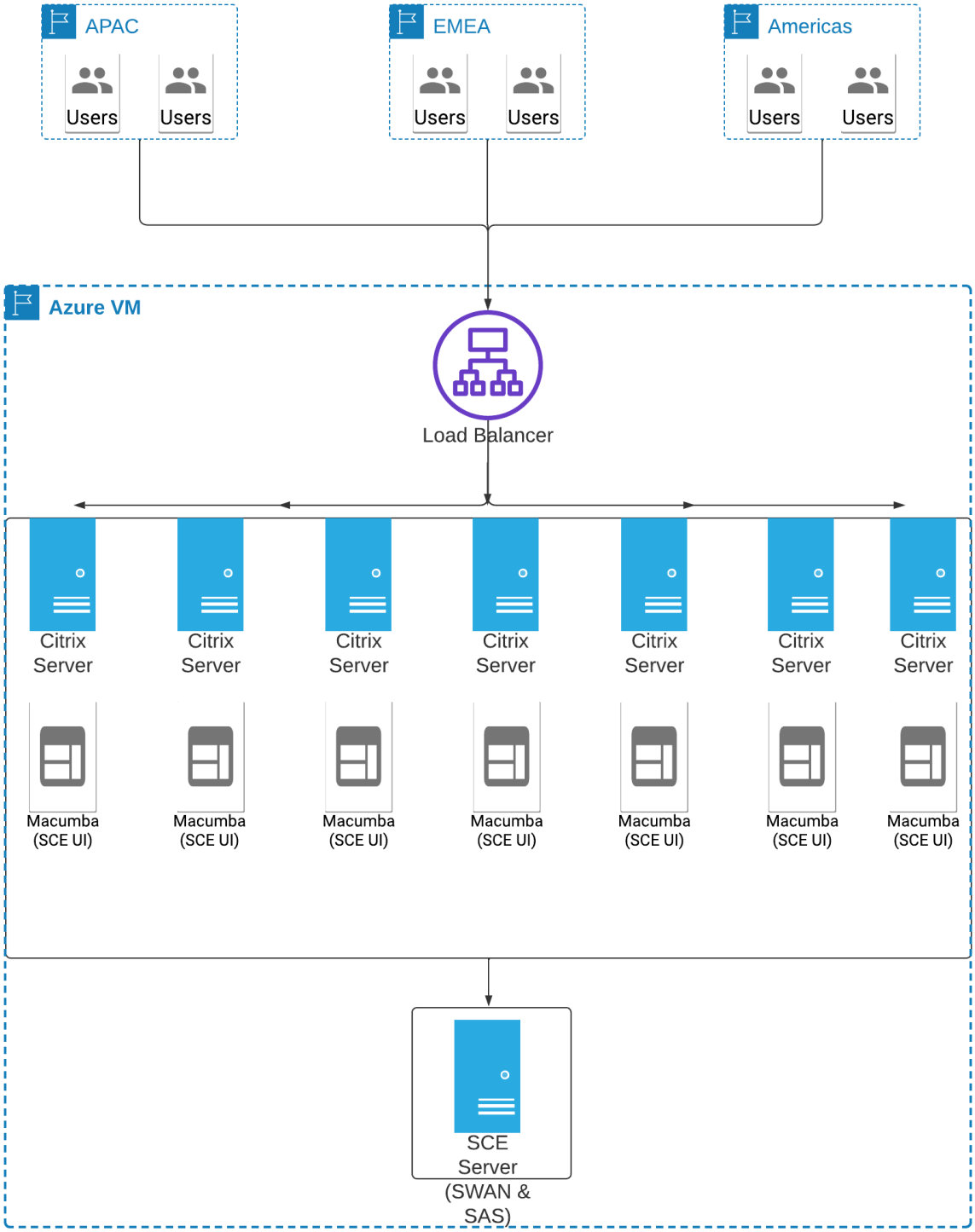
Code Optimization and Documentation
The refactoring process involved significant improvements to the Macumba application, particularly in areas such as auto-save functionality and startup efficiency. By reverse-engineering the source code, we partnered with a service provider (TSRI) and generated comprehensive documentation, supported by a large language model (LLM)-powered chatbot, to facilitate better understanding and maintenance of the system.

Infrastructure and Automation Enhancements
Our migration strategy included the implementation of Infrastructure-as-Code, enabling the seamless provisioning of new environments with Linux and SAS. This approach was complemented by establishing robust CI/CD pipelines for SWAN and Macumba, ensuring rapid and reliable software deployments. Additionally, we developed extensive test suites for system, functional, and load testing, enhancing the overall reliability and performance of the applications.

User Experience and Observability
To improve user experience, we replaced the legacy Citrix environment with AWS AppStream, which provides stable session management regardless of user load. This change significantly enhanced the usability and accessibility of our applications. Furthermore, we integrated AWS CloudWatch for enhanced observability, allowing for improved monitoring and support through detailed log analysis.

Foundation for Future Innovation
In collaboration with the Data Science Ecosystem (DSE) team, we are establishing a foundation for open-source capabilities using Posit R Studio. This initiative is part of our broader strategy to foster innovation and flexibility in our analytical processes.

Architecture of the legacy SCE



This is a **traditional, centralized virtualization architecture** using Citrix for application delivery, hosted on Azure infrastructure.

Global User Base

APAC (Asia-Pacific): Multiple user groups
EMEA (Europe, Middle East, Africa): Multiple user groups
Americas: Multiple user groups

Characteristics:

Globally distributed user population
All regions connect to a single, centralized Azure environment via a Citrix Gateway

Load Balancing

Single Load Balancer: Distributes user sessions across Citrix servers
Purpose: Traffic distribution and basic high availability

Application Delivery Layer - Citrix Farm

Citrix Infrastructure

Multiple Citrix Servers: Horizontal scaling for user sessions
Citrix Virtual Apps/Desktops: Application virtualization and delivery
Session-based Architecture: Each user gets a Citrix session on one of the servers

Application Deployment

Macumba (SCE UI): Pharmaceutical analytical interface deployed on every Citrix server
Uniform Configuration: Same application stack on all servers
Resource Pooling: Users can connect to any available server

Backend Processing - SCE Server

Centralized Processing Server

SCE Server (SWAN & SAS):
SWAN: Software Wide Area Network connectivity
SAS: Statistical Analysis System - the core analytics engine
Shared Resource: All Citrix servers connect to this single backend
Processing Architecture: Citrix servers act as presentation layer; SCE server handles computation

Architecture of the Refactored SCE

This architecture diagram illustrates a **modernized, cloud-native architecture to deploy the refactored legacy SCE**.

AWS AppStream 2.0 Infrastructure

- **Fleet - Streaming Instances**: Auto-scaled, on-demand virtual application servers
- **Stacks**: Configuration templates for streaming sessions
- **Images**: Golden images containing pre-configured applications

SCE Processing VPC (Core Analytics Environment)

- **Region**: Frankfurt
- **Multi-AZ Design**: Dual availability zones for high availability
- **Compute Tier**:
  - **SAS Server**:
    - SAS 9.4 Compute engine
    - Samba file sharing for Windows compatibility
    - **Work Storage**: Dedicated storage volumes for active processing
- **Standby Availability Zone (mirror configuration)**
  - Identical architecture to Primary AZ
  - **SAS Server**: SAS 9.4 Compute + Samba
  - **Work Storage**: Replicated storage

AWS AMI Images

- **Custom Machine Images**: Pre-configured SAS and application servers
- **Version Control**: Standardized deployments across AZs
- **Rapid Scaling**: Quick instance provisioning

Supporting Infrastructure

- **CloudWatch Logs**: Centralized log aggregation and analysis

Advantages of the modernized architecture

Aspect	Legacy (Citrix/Azure VM)	Modern (AppStream/AWS)
Scaling	Manual, fixed capacity	Auto-scaling, elastic
Availability	Single region, limited HA	Multi-AZ, built-in redundancy
User Experience	Citrix protocol, latency-sensitive	Streaming, adaptive to bandwidth
Cost Model	Fixed costs (always-on VMs)	Pay-per-use (streaming minutes)
Deployment	Manual server provisioning	AMI/Image-based, automated
Geographic Reach	Centralized, potential latency	Regional deployment options
Disaster Recovery	Not visible/limited	Multi-AZ with automated failover
Maintenance	Updates across multiple servers	Golden image updates

In summary, the legacy architecture represented a **traditional enterprise virtualization approach** that served its purpose but hit limitations in scalability, cost-efficiency, and operational agility that the modernized cloud-native architecture addresses.

Conclusion

The successful migration and refactoring of SWAN and Macumba mark a significant milestone in Bayer's journey to modernize its Statistical Computing Environment. By addressing the challenges outlined in our problem statement, we envision to achieve potential cost savings and operational efficiencies. Our efforts not only preserved the core business functionalities but also positions Bayer to integrate advanced components like SAS Viya and POSIT, ensuring a resilient and scalable SCE for the future while underscoring our commitment to leveraging cutting-edge technologies to enhance our analytical capabilities.

CONTACT INFORMATION

Authors
Author Name: Vaithi Bharath, Associate Director, R&D Data Science & AI
Company: Bayer
Email: vaithi.bharath@bayer.com

Author Name: Srinath Viswanathan, Software Engineering Lead, IT
Company: Bayer
Email: srinath.viswanathan@bayer.com

Author Name: Srinivas Veeragoni, Senior Director, R&D Data Science & AI
Company: Bayer
Email: srinivas.veeragoni@bayer.com

Refactoring & Modernization Service Provider
Company: The Software Revolution Inc. (TSRI)
Address: 11332 NE 122nd Way #300, Kirkland, WA 98034
Contact: Thomas Kucher
Phone: +1-617-901-9148
Email: tkucher@tsri.com
Website: https://tsri.com/

About TSRI
Since 1995, TSRI has offered a truly unique, language- and industry-neutral automated legacy software modernization service for government and commercial customers worldwide. Our staff includes scientists at the forefront of artificial intelligence-based application transformation, and our facility and specialized equipment is capable of handling any system modernization project (including classified work) regardless of language, size, or complexity.