Цифровой полигон Высшей инжиниринговой школы МИФИ для инфраструктурного обеспечения образовательных и учебно-практических проектов

Г.Г. Гаджилов, А.В. Хатунов, Т.А. Волошин, М.Г. Жабицкий

Аннотация - Статья описывает концепцию, архитектуру и результаты опытной эксплуатации «Цифрового полигона» ВИШ НИЯУ МИФИ — цифровой инфраструктуры для объединения образования, НИР и прикладных проектов с индустриальными партнёрами. Показано, как разрыв ожиданий между академической и промышленной средами (в терминах уровней технологической готовности, TRL) трансформируется в требования к сервисам, политике доступа и наблюдаемости. Предложена целевая архитектура, основанная на виртуализации (Proxmox VE), инфраструктуре как коде (Ansible) и едином контуре наблюдаемости (Prometheus/Grafana, распределённый трейсинг), а также регламенты эксплуатации для учебно-исследовательских и пилотных производственных сценариев. В рамках кейсов продемонстрированы: локальный инференс больших языковых моделей на рабочих станциях с несколькими GPU и стенд микросервисного приложения (≈15 сервисов) с трассировкой и бизнес-метриками. По итогам внедрения зафиксированы улучшения эксплуатационных показателей: рост доступности до ~99,5%, снижение числа инцидентов на ~78%, сокращение времени развёртывания типовых служб до 1–2 часов; для микросервисного стенда — уменьшение времени обнаружения и устранения отказов (TTD/TTR) на 67% и 58% соответственно. Научная и практическая новизна работы — в интеграции TRL-подхода к взаимодействию с индустрией с воспроизводимым инженерным шаблоном кампусной платформы (виртуализация + IaC + наблюдаемость) и показом её применимости для on-prem задач ИИ. Намечены шаги дальнейшего развития: кластеризация, унификация IAM/SSO и политика резервного копирования.

Ключевые слова — цифровой полигон, виртуализация, мониторинг, инфраструктура как код, наблюдаемость, микросервисная архитектура, масштабируемость, отказоустойчивость.

Введение

На текущем уровне зрелости научно-технологической системы современного общества в России и в мире можно выделить несколько типов центров научно технологического развития:

- научно-исследовательские университеты;
- R&D подразделения технологических компаний и корпораций;
- инициативные стартапы;
- независимые научно-исследовательские институты.

В рамках данной работы мы сосредоточимся на анализе специфических роли и места научно-исследовательского университета как одного из центров экосистемы технологического развития и того инфраструктурного инструментария, который может обеспечивать эту центральную роль передовых университетов для исследований и разработок в сфере информационных технологий. Мы опишем опыт инновационного структурного научно образовательного подразделения одного из ведущих российских технологических университетов – Высшей инжиниринговой школы (ВИШ) НИЯУ МИФИ в проектной реализации такого инфраструктурного ядра такой экосистемы. Подход аналогичен логике работы [1]. В работе представлена методология проектирование и создания такого объекта – Цифрового полигона ВИШ, описана архитектура и ключевые компоненты системы, сформулирована гипотеза о вместе и роли системы в образовательном процессе и в генерации научно-исследовательской работы в университете, в интересах профильных индустриальных партнеров (на примере предприятий ГК Росатом), а также технологических стартапов студентов, аспирантов и молодых ученых.

I. Постановка задачи

Высшая инжиниринговая школа НИЯУ МИФИ - структурное подразделение университета, созданное в 2017 году для подготовки кадров высшей квалификации (магистратура) в интересах предприятий высокотехнологичной индустрии, прежде всего входящих в структуру ГК Росатом. Образовательные программы ВИШ ориентированы на подготовку специалистов по цифровой трансформации высокотехнологичных производственных предприятий и разработки передовых цифровых продуктов в соответствии с парадигмой Четвертой промышленной революции. Образовательный процесс ориентирован на сетевое взаимодействие с предприятиями атомной и других передовых отраслей российской экономики С одним из ведущих университетов Российской Федерации – НИЯУ МИФИ. Образовательными направлениями подготовки магистратуры ВИШ являются «Информационные системы и

технологии» (основной род деятельности выпускников после окончания - разработка информационных систем и инновационных цифровых продуктов) и «Системный анализ и системная инженерия» (основной род деятельности выпускников после окончания - аналитическая работа в ходе всего жизненного цикла информационных систем и инновационных цифровых продуктов).

В структуре образовательного процесса ВИШ можно выделить четыре основных формата образовательной деятельности для студентов. Во-первых, это классическая системная образовательная составляющая – лекции, семинары, цифровые и очные лабораторные работы по фундаментальным и прикладным предметам учебной программы направлений подготовки. Во-вторых, это погружение в сетевую компоненту, направленную на обучение в соответствии с углубленной специализацией на специфических знаниях, умениях и навыках, востребованных в ходе цифровой трансформации и цифрового развития атомной отрасли.

Далее, крайне важной составляющей является групповая проектная работа студентов. В данном формате крайне заинтересованы Индустриальные партнеры, которые формулируют постановку проектных задач, и имеют возможность мониторить ход выполнения через наставников проекта, выделенных от заинтересованной организации. Это дает возможность отсмотреть будущих претендентов на рабочие места не только по их знаниям и компетенциям, но и по отношению к взаимодействию с потенциальным работодателем, знаниям и навыкам работы с конкретным стеком технологий, готовности и умения встроиться в обычаи делового оборота компании-работодателя и практической пользе при реализации проекта. Фактически, такой подход дает возможность «почти реальной» и достаточно длительной стажировки в ходе всех двух лет обучения в магистратуре и поэтапного вовлечения в команду организации.

Наконец, последним форматом является НИР студентов, переходящие в исследование в рамках выпускной квалификационной работы (ВКР). Этот формат достаточно обычен на первый взгляд, однако в случае увязки с реализацией учебно-практического проекта и взаимодействия со специалистами Индустриального партнера как консультантами ВКР он крайне интересен для обеспечения бесшовной интеграция в команду предприятия на конкретную кадровую позицию с готовым заделом по направлению работы организации. При этом издержки на адресную подготовку и упреждающую интеграцию потенциальных сотрудников для работодатели существенно снижены по сравнению с набором молодых выпускников ВУЗов через открытый рынок. Равно как и риски несоответствия взаимных ожиданий претендентов и работодателей, выявляющихся в первые периоды работы даже при условии испытательного срока. Также в режиме углубленных НИР магистрантов во взаимодействии с индустриальными партнерами заинтересован университет, как с точки зрения выявления перспективных научно-педагогических кадров, в дальнейшем вовлекаемых через механизмы аспирантуры и преподавательской деятельности по совместительству, так и для развития углубленного научно-технического взаимодействия с Индустриальным партнером.

Для целей настоящей работы важно охарактеризовать основной инструментарий, используемый студентами, преподавателями и консультантами со стороны индустриальных партнеров в образовательном процессе. Естественно, это связано со специальностями обучения. Поскольку цель ВИШ - цифровые технологии, то задача их реализации и развертывания на полигоне достаточной мощности является непременным условием м для практической компоненты подготовки, и для накопления результатов интеллектуальной деятельности, полученных во всех формах образовательного процесса. Чрезвычайно быстрая динамика развитие инструментов цифровой трансформации всех сфер деятельности приводит к тому, что решаемые в образовательном процессе задачи меняются крайне динамично. В реальности уровень обновления задач всех видов для самостоятельной работы студентов во всех образовательных процессов усредненно составляет 30-50% год к году. Частным следствием является то, что накопления опыта и знаний реализуется не через традиционные формы учебников и задачников, а в современном цифровом формате GIT-инструментов, динамических баз знаний, библиотек разработанных магистрантами, аспирантами и сотрудниками ВИШ цифровых решений, баз данных и дата сетов цифровых решений.

Наконец, остановимся на специфике взаимодействия современного университета с индустриальными партнерами. Для анализа характера такого взаимодействия необходимо понимание современных методов развития технологии и продуктов в высокотехнологичной сфере. Общепринятым методом анализа является классификация жизненного цикла научно-техническая разработки по уровню технологической готовности (Technology Readiness Level, TRL). Для определенности будем оперировать данным понятием в терминологии нормативной документации ГК Росатом [2,3]. И в этой точке рассуждений мы можем диагностировать явные противоречия между подходами. Они представлены в Таблице 1.

Конфликт философий

Сущность	Университет	Индустриальный партнер
Целевой результат	Образованные свободные люди	Востребованный экономикой товарный продукт
Ошибки и несоответствия	В образовании и науке нормальны	Однозначный негатив
Способ деятельности	Вузовский исследователь обычно занимается тем, что интересно. Часто всю жизнь	БОльшая часть менеджеров и производственников выполняют поставленные задачи.
Уровень связи и иерархии	Слабые связи в системе. Иерархия строится на авторитете.	Жесткие связи. Иерархия строится на структуре и дисциплине.
Уровень TRL	Результаты интеллектуальной деятельности (РИД) – TRL 1-2-3-4-5, редко выше	Заинтересованы продуктами высоких уровней TRL 6-7-8-9
Доступ к информации	Намеренно свободный, открытые публикации	Режимы защиты информации. Защита "KNOW-HOW", коммерческая тайна, секреты производства.

Таблица 1. Анализ подходов ВУЗа и ИП.

Детальный анализ обоснований и последствий указанных противоречий интересен, однако выходит за пределы настоящей статьи. Но абсолютно понятно, что для преодоления явно видимого разрыва необходимо общее пространство. Оно необходимо и в форматах деятельности - и мы показали выше его структуру, способы формирования и порядок взаимодействия в нем. Но также оно необходимо и в части инфраструктуры и организации ее эксплуатации. Инфраструктура индустриального партнера не предназначена для некоммерческой, экономически неэффективной деятельности и практически всегда имеет серьёзные ограничения с точки зрения доступа и информационной безопасности, особенно для реализации поисковых проектов, особенно людьми не имеющими формализованных жестких обязательств перед предприятием. При этом, такие проекты могут быть перспективны и с точки зрения развития компетенций, в том числе для сотрудников индустриальных предприятий, и с точки зрения начальных стадий прототипирования приложений прорывных технологий «на через шаг». Исключение таких форматов из сферы деятельности означает консервацию технологического уровня и часто - остановку развития.

В Высшей инжиниринговой школе НИЯУ МИФИ данная проблема разрешается через развитие специализированной инфраструктурной сущности - Цифрового полигона ВИШ. Отдельным аспектам ее проектирования, развертывания и функционирования посвящена настоящая работа.

II. Функциональные и нефункциональные требования к цифровому полигону

Мы можем сформулировать основных стейкхолдеров функционирования цифрового полигона ВИШ их основные потребности. Аналогами можно рассматривать [5-7]. В качестве стейкхолдеров можем рассматривать участников классическая компоненты образовательной деятельности - преподавателей и студентов магистратуры. Их потребности:

- Функцию локального и удаленного доступа к ресурсам и контенту цифрового полигона.
- Управляемый доступ к структурированному в соответствии с учебными программами и дисциплинами образовательному контенту (лекционный материал, задание для самостоятельного выполнения, методические материалы и руководств);
- Доступ к инструментам (программным средам, инструментом разработки, специализированным цифровым продуктом) для использования в ходе образовательного процесса. На современном уровне сюда входят инструменты работы системами искусственного интеллекта, включая доступ к облачным и контурным большим лингвистическим моделям с возможностью накапливать промежуточные и итоговые результаты диалоговых сессий.
- Простые процедуры сдачи результатов в ходе занятий и по результатам выполнение домашних заданий, курсовых проектов и других результатов самостоятельная работа студентов.
- Возможность для преподавателей структурирование и накопления результатов работы студентов, включая настраивая функции проверки цифровых реализаций заданий для самостоятельной работы.

Второй важной группой стейкхолдеров является проектные объединения - включая группы по учебно-практическим проектам, выполняемым под руководством наставников со стороны индустриальных партнеров, а также научно-исследовательские группы, решающие задачи в рамках фундаментальных и прикладных исследований в инициативном порядке, в рамках различных форм различных форм поддержки научно-исследовательской деятельности, а также в рамках договорной научной работы по заказу предприятий и организаций. Для таких групп необходимо обеспечить выполнение следующих функций:

- Развертывание единого информационного пространства проекта.
- Структурирование информационного пространства проекта в зависимости от его потребностей.

- Развертывание необходимых для данного проекта цифровых инструментов (сред разработки, специализированных инструментов цифровой инженерной деятельности, например таких как BIM-инструменты, системы виртуальной реальности, системы искусственного интеллекта и других).
- Доступ к инструментам масштабирования разрабатываемых информационных систем, прототипов цифровых продуктов и баз данных.
- Функция надежного контроля доступа в информационное пространство проектной группы.
- Функцию локального и удаленного доступа к цифровому полигону,
- Функцию управляемого и прогнозируемого выделения вычислительных ресурсов на периоды интенсивной работы проектные группы либо пиковых нагрузок на разрабатываемые системы.
- Функцию динамического свертывания и активации используемых проектной группой цифровых инструментов в целях эффективного управление ресурсами для конкурирующих групп пользователей.
- Функцию архивирование и надежного хранения архивов, а также их развертывания для продолжения деятельности.
- Функцию мониторинга состояния функционирование цифровых продуктов как общими инструментами цифрового полигона, так и специально развернутыми в рамках информационного пространства проектной группы специфическими инструментами.

Далее, обратим внимание на специфику использования информационных ресурсах в рамках цифрового полигона образовательной организации. Задачи, решаемые основным производственными подразделениями (кафедрами, факультетами, лабораториями) имеют одновременно переменчивый и инновационный характер, в значительной части случаев не превращаясь в продукт - поскольку пользователи (и студенты, и преподаватели) не продают его, а используют для развития - учебных задач, исследований, испытаний, прототипов. При этом требуется частая и быстрая смена форматов - новые курсы, новые проекты, подключение и последующее отключение прав доступа для большого количества пользователей, перед которыми у информационной площадки ограниченная ответственность. А ошибка разработчиков (студентов и действующих ученых) ведет к полезному эффект увеличения знаний, а не к потери платежеспособным спроса. Соревновательные мероприятия в сфере информационных технологий, задача заранее неизвестным решением, тренировки хакеров - все это нормальное явление для университета. Чтобы получить серьезные навыки нужно знать и уметь не только «как можно», но и получить возможность попробовать «как нельзя». Но, дополнительно к этому необходимо приводить инфраструктуру и ландшафт система обратно к норме. И здесь, именно в университете, возникает еще один класс стейкхолдеров - администраторы и архитекторы цифрового полигона, имеющие свои отдельные и необычные для стандартного рынка информационных технологий потребности. Это связано с необходимостью постоянной смены и перестройки информационного ландшафта. Администраторы цифровых полигонов университетов вынуждены обеспечивать условия работы для самых неожиданных и экзотических стеков технологий - и в силу разнообразия образовательных курсов и практик, и в силу непрерывного, регулярного потока инновационных решений - особенно в современных условиях и при быстро развивающихся технологиях. Внутренняя потребность у этого классов стейкхолдеров – обеспечить высокую производительность труда в администрировании ресурсов цифрового полигона, связанная с большим количеством пользователей, разнообразием их запросов как по технологическому стеку, так и по потребляемым ресурсам и масштабу задач, а также большому объему деятельности по восстановлению, в том числе и в результате работы с необычным, а иногда и неудачными или даже разрушительными решениями, рождающимися в ходе образовательного процесса и поисковый научной деятельности.

В результате можно сделать вывод, что цифровой полигон профильного ИТ подразделения современного университета отличается от стандартных замкнутых контуров разработки цифровых продуктов фирм-производителей, и от инфраструктурных проектов центров обработки данных для предоставления сервисов и инфраструктуры для продуктивной эксплуатации готовых информационных продуктов.

Сформулируем также нефункциональные требования для описываемого нами нетривиального инфраструктурного объекта – цифрового полигона образовательно-исследовательского подразделения одного из ведущих университетов, тесно взаимодействующего с индустриальными партнерами в сфере разработки современных передовых информационных продуктов. Итак, необходимо обеспечить:

- широкий технологический стек с введением в него необычных и нетривиальных технологий, а также сочетаний цифровых продуктов и технологий;
- необходимо обеспечить информационную защищённость для каждого проекта;
- необходимо обеспечить возможность быстрой управляемый смены ландшафта полигона в соответствии с потребностями разнообразного

образовательного процесса и исследовательской деятельности;

- необходимо обеспечить регулярное надежное сохранение данных и информационных структур цифрового полигона в связи с возможными угрозами, в том числе в ходе выполнения исследований и экспериментов разработчиками с недостаточным опытом, а иногда и ответственностью (студенты бывают разные);
- необходимо обеспечить накопление результатов регулярной интеллектуальной деятельности большого количества разработчиков с учетом авторства каждого из них, в том числе для случаев недостаточной ответственности авторов;
- необходимо обеспечить быстрое восстановление как всего ландшафта цифрового полигона, так и его отдельных компонент и структурированных архивов;
- необходимо обеспечить интенсивный процесс постоянного ввода новых пользователей, и блокировку прав выпускников
- необходимо обеспечить эффективное и прозрачное управление доступом к ресурсам для большого количества пользователей, суммарные потребности которых априори превышают в сумме имеющиеся в распоряжении администратора с учетом приоритизации по внешним факторам (регулярные образовательные мероприятия имеют приоритет перед инициативными разработками, однако при этом должно обеспечено динамическое управление ресурсов с учетом реальных потребностей и расписания).

III. Функционирование цифрового полигона на этапе прорывного развития систем искусственного интеллекта

Взрывное развитие систем искусственного интеллекта в сфере информационных технологий существенно меняет весь подход к деятельности профильного специалиста [8]. С момента проектирования и начала развертывания цифрового полигона ВИШ принципиально возросла роль и значимость компонент, связанных с разными реализациями таких инструментов. Это включает в себя и уже почти традиционные на сегодняшний день нейронные сети с технологией глубокого машинного обучения, и неожиданно возникший как чудо для широкого круга пользователей феномен больших лингвистических моделей (LLM), внедрение которых во все виды интеллектуальной деятельности является основным революционным событием для последних 3 лет. При этом мы прогнозируем интенсивное использование как облачных реализации LLM в доступных для их разработчиков открытых системах, так и решение задачи ограничения доступа к результатом работы с интеллектуальными ассистентами путем развертывания на базе собственной информационной инфраструктуры больших и малых предприятий закрытых контурных моделей для решения задач, содержащих технологическую, коммерческую и иную информацию ограниченного доступа. Естественно, в состав цифрового полигона ВИШ включаются компоненты, обеспечивающие построение, обучение и использование классических нейросетей, а также развертывание больших лингвистических моделей в замкнутом контуре и доступ к облачным LLM для выполнения работ с наиболее актуальным на сегодняшний день цифровым инструментом. В этой сфере крайне уместного оказалось использование сходной инфраструктуры различными по содержанию и назначению цифровыми инструментами. Так, графические ускорители оказались широко востребованными для совершенно разных технологических областей, входящих в образовательной набор ВИШ НИЯУ МИФИ. В рамках образовательной программы и проектной деятельности магистрантов и аспирантов ВИШ реализуется совершенно независимо предметы «Цифровое проектирование сложных инженерных объектов», «Инженерная виртуальная реальность» и целый букет специальных образовательных курсов по различным технологиям и видам приложений систем искусственного интеллекта. С точки зрения используемой аппаратной части все эти разные по сферам приложений задачи основаны на специализированных вычислительных мощностях графических ускорителей. Оборудование цифровых учебных классов и лабораторий велось техникой соответствующих конфигураций, и в результате на текущий момент мы имеем достаточно большой набор рабочих станций, пригодных для выполнения задач с использованием технологии искусственного интеллекта. Соответственно, одним из функциональных модулей цифрового полигона ВИШ является система организации доступа и оркестрации использования компьютерный техники для задач обучение нейросетей и других приложений ИИ. Это же компьютерное оборудование продолжает интенсивно использоваться в специализированных курсах и учебно-практических проектах трехмерного цифрового проектирования разработки системные приложения виртуальной реальности.

Этот аспект единой инфраструктурной базы для подготовки специалистов и выполнение исследований для различных специализаций сфере цифровизации является важной иллюстрацией целесообразности и эффективности стратегического проектирование Цифрового полигона университета с последующим системным развертыванием, развитием и эксплуатацией для обеспечения качественной подготовки специалистов и выполнения на современном уровне научно-

исследовательских работ для подразделений ведущих университетов, в особенности вовлеченных в сетевые образовательные программы инженерно-технологические разработки в сфере информационных цифровых технологий С высокотехнологичными инновационными предприятиями.

IV. Примеры исследований и разработок, выполненных магистрами и аспирантами ВИШ в интересах создания и развития Цифрового полигона

IV.I Разработка и интеграция сервиса доступа к большим языковым моделям на локальных вычислительных ресурсах Цифрового полигона ВИШ

В рамках исследований по оптимизации запуска крупных языковых моделей (LLM) на локальных вычислительных ресурсах был развёрнут стенд, ориентированный на локальное взаимодействие с пользователями. В качестве пользовательского интерфейса на стенде установлено приложение LM Studio, обеспечивающее доступ к LLM в удобной графической оболочке без необходимости использования командной строки. Это позволяет пользователям формулировать запросы, получать ответы и проводить эксперименты с различными моделями прямо на рабочей станции, что особенно важно в условиях ограниченного доступа к облачным сервисам. Конфигурация стенда:

- Процессор: Intel Core i7 13700KF
- Оперативная память: 64 ГБ DDR4
- Графический ускоритель: NVIDIA RTX 3060 12 ГБ
- ОС: Windows 11

В рамках работы по оптимизации инфраструктуры и исследованию масштабируемости инференса LLM к стенду была добавлена вторая видеокарта NVIDIA RTX 3060. Это позволило протестировать возможности многокарточного режима работы. Для экспериментов использовалась библиотека llama.cpp, поддерживающая запуск квантованных моделей в формате GGUF с автоматическим распределением слоёв между несколькими GPU. Библиотека использует CUDA и cuBLAS для выполнения операций линейной алгебры на GPU и обеспечивает стабильную работу в многокарточной конфигурации. Измерения производительности проводились с помощью утилиты llama-bench, которая позволила сравнить скорость обработки промпта и генерации новых токенов в двух режимах: при использовании одной и двух видеокарт.

В тестировании участвовали три модели различных размеров:

Qwen 3 32B (примерно 32 миллиарда параметров, квантована до Q4_K_M, размер 18.4 ГБ)

Mistral Small 3.1 24B (24 миллиарда параметров, Q4_K_M, 13.3 ГБ)

Llama 2 7B (7 миллиардов параметров, Q8_0, 6.67 ГБ)

Все модели были запущены в режиме инференса на заранее подготовленных промптах длиной 128 и 512 токенов. Замерялась как скорость чтения (prompt processing), так и скорость генерации новых токенов (text generation), единицы измерения – токены в секунду (t/s). Запуск на двух видеокартах осуществлялся с автоматическим распределением слоёв модели между GPU средствами llama. В таблице 1 приведены результаты тестов.

	Модель	**1 GPU**	**2 GPU**
Чтение промпта, t/s	Qwen 3 32B	45.41 ± 0.23	434.79 ± 12.84
	Mistral Small 3.1 24B	81.26 ± 1.74	673.60 ± 2.77
	Llama 2 7B	1969.34 ± 22.87	1965.21 ± 29.53
Генерация текста, t/s	Qwen 3 32B	1.11 ± 0.01	15.12 ± 0.26
	Mistral Small 3.1 24B	3.55 ± 0.05	21.58 ± 0.10
	Llama 2 7B	41.77 ± 0.55	41.05 ± 0.17

Таблица 2 – Результаты тестов языковых моделей для различных конфигураций рабочих станций

В ходе тестирования было установлено, что прирост производительности при добавлении второй видеокарты напрямую зависит от объёма модели по отношению к доступной видеопамяти одной карты. Модели меньшего размера, такие как Llama 2 7B, полностью помещаются в 12 ГБ VRAM одной RTX 3060, что позволяет им работать исключительно в пределах видеопамяти без необходимости обращения к оперативной памяти. В этом случае задействование второй видеокарты не даёт заметного прироста, так как вычислительная нагрузка укладывается в возможности одного ускорителя. Иная ситуация наблюдается с моделями объёмом 24B и 32B параметров, они не умещаются в память одной RTX 3060. Тогда при запуске с одной видеокартой часть слоёв модели вынужденно размещается в оперативной памяти, и при инференсе происходит постоянный обмен между GPU и CPU через шину PCIe, что значительно снижает скорость генерации. При подключении второй видеокарты библиотека llama.cpp автоматически распределяет слои между двумя GPU, и каждая карта обрабатывает только те части модели, которые умещаются в её собственной памяти. Это позволяет избежать обращения к медленной системной памяти и полностью задействовать высокоскоростной доступ к VRAM обеих карт, что приводит к значительному ускорению как обработки

промпта, так и генерации текста.

Полученные результаты подтвердили эффективность масштабирования в пределах одного узла при наличии нескольких GPU, что позволяет существенно ускорить инференс крупных моделей без перехода к распределённой среде. Следующим шагом в развитии стенда станет переход к многомашинной архитектуре с объединением нескольких вычислительных узлов по сети. Планируется исследовать два направления: запуск отдельных экземпляров модели на каждом узле с использованием балансировщика запросов, а также распределение модели между узлами с разбиением параметров и межмашинным взаимодействием. Первый подход позволяет эффективно масштабировать систему, не требует сложной сетевой инфраструктуры и обеспечивает устойчивость к сбоям отдельных узлов, в то время как второй ближе к промышленным распределённым системам, но он требует предварительной подготовки сетевой и программной инфраструктуры.

IV.II Виртуализированая инфраструктура цифрового полигона ВИШ для размещения ключевых сервисов

Анализ ограничений существующей монолитной архитектуры и требований к цифровому полигону ВИШ обосновывает необходимость перехода к разнесенной архитектуре с изолированным размещением сервисов. Основным принципом разнесенной архитектуры является разделение функциональности системы на независимые компоненты, каждый из которых выполняется в собственной изолированной среде. В контексте виртуализированной инфраструктуры это означает размещение каждого сервиса в отдельной виртуальной машине или контейнере.

Преимущества разнесенной архитектуры включают устранение единой точки отказа, поскольку сбой одного компонента не влияет на работу других сервисов. Изоляция ресурсов позволяет предотвратить ситуации, когда высокая нагрузка на один сервис влияет на производительность других компонентов. Независимое масштабирование компонентов обеспечивает возможность выделения дополнительных ресурсов только тем сервисам, которые испытывают повышенную нагрузку, без необходимости масштабирования всей системы. Это особенно важно в образовательной среде, где различные сервисы могут иметь различные паттерны использования. Упрощение обслуживания достигается за счет возможности обновления, перезапуска или изменения конфигурации отдельных компонентов без влияния на работу других сервисов. Это критически важно для минимизации времени простоя в образовательной среде.

Технологическая гибкость обеспечивается возможностью использования различных операционных систем, версий программного обеспечения и конфигураций для различных компонентов в зависимости от их специфических требований. Принципиальная схема цифрового полигона ВИШ представлена на Рисунке 1, а схема взаимодействия - на рисунке 2.

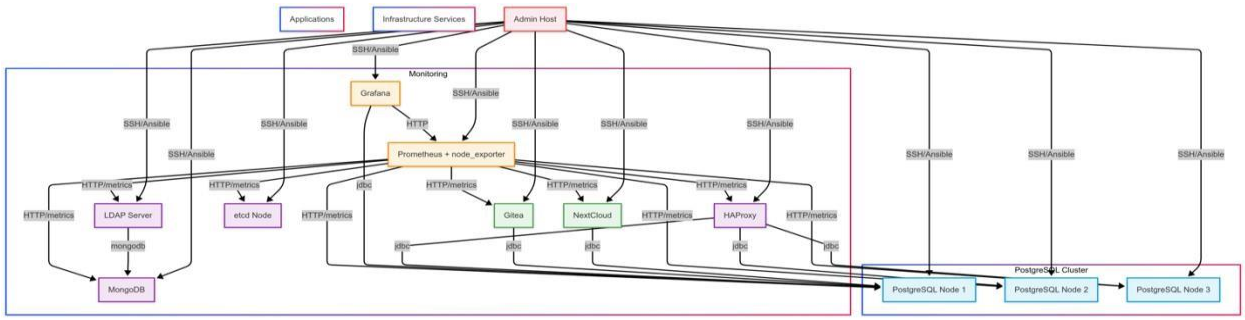

Рисунок 3 – Принципиальная схема Полигона ВИШ

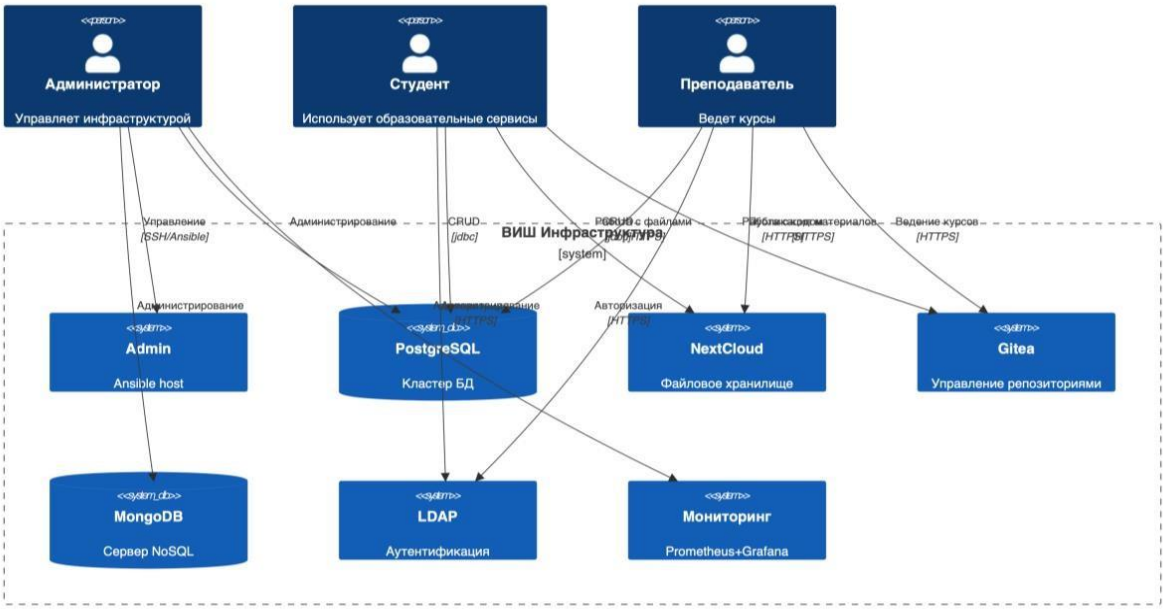

Рисунок 2 – Схема взаимодействия

Проведенный анализ результатов реализации и внедрения виртуализированной инфраструктуры цифрового полигона ВИШ НИЯУ МИФИ позволяет объективно оценить степень соответствия достигнутых результатов изначально поставленным целям и задачам проекта.

Главная цель проекта заключалась в создании отказоустойчивой, масштабируемой и централизованно управляемой ИТ-инфраструктуры цифрового полигона на базе виртуализированной среды. Эта цель полностью достигнута. Разработанная архитектура с разнесенным размещением сервисов в отдельных виртуальных машинах обеспечила требуемую изоляцию компонентов и устранила единую точку отказа, характерную для предыдущей монолитной архитектуры.

Задача обеспечения отказоустойчивости решена через изоляцию сервисов, автоматическое восстановление при сбоях и комплексную систему мониторинга. Тестирование подтвердило, что сбой отдельного компонента не приводит к каскадному отказу всей системы, а среднее время восстановления соответствует или превосходит установленные требования.

Задача обеспечения масштабируемости решена через выбор архитектуры, которая поддерживает как вертикальное, так и горизонтальное масштабирование. Виртуализация на базе Proxmox VE обеспечивает гибкие возможности по выделению ресурсов и добавлению новых компонентов без кардинальных изменений в архитектуре.

Задача централизованного конфигурирования успешно решена с использованием Ansible для автоматизации процессов настройки и сопровождения инфраструктуры. Созданная система плейбуков и ролей обеспечивает единообразие настроек, минимизирует человеческий фактор и значительно сокращает время, необходимое для типовых операций.

Задача организации комплексного мониторинга решена через внедрение стека Prometheus/Grafana, который обеспечивает сбор и визуализацию метрик со всех компонентов инфраструктуры. Созданные дашборды и система алертинга обеспечивают проактивное выявление проблем до их влияния на пользователей.

Образовательные цели проекта, связанные с демонстрацией современных DevOps практик и предоставлением студентам актуальных инструментов разработки, также полностью достигнуты. Инфраструктура успешно интегрирована в образовательный процесс и получила положительные отзывы от студентов и преподавателей.

В ходе реализации проекта был достигнут ряд значимых технических и образовательных результатов, которые выходят за рамки первоначально поставленных задач и создают дополнительную ценность для цифрового полигона ВИШ.

Повышение надежности инфраструктуры является одним из ключевых достижений проекта. Доступность сервисов увеличилась с 95% в предыдущей архитектуре до 99.5% в новой системе. Количество инцидентов, требующих вмешательства администраторов, сократилось на 78%, что свидетельствует о высокой стабильности созданной инфраструктуры.

Оптимизация использования ресурсов достигнута за счет гибкого распределения вычислительных мощностей между

виртуальными машинами. Анализ использования ресурсов показал, что средняя утилизация CPU увеличилась с 25% до 60%, а использование RAM с 40% до 75%, что свидетельствует о более эффективном использовании аппаратных ресурсов.

Автоматизация рутинных операций привела к сокращению времени, необходимого для выполнения типовых задач. Время развертывания нового сервиса сократилось с 2-3 дней до 1-2 часов, а время обновления системы с 8 часов до 30 минут. Эти улучшения обеспечивают более гибкое реагирование на потребности образовательного процесса.

Интеграция современных практик DevOps в образовательный процесс позволила студентам получить практический опыт работы с промышленными инструментами и методологиями. По результатам опроса, 87% студентов отметили, что работа с новой инфраструктурой повысила их профессиональные компетенции и подготовила к требованиям рынка труда.

Создание комплексной системы мониторинга обеспечило беспрецедентную видимость всех аспектов работы инфраструктуры. Это не только улучшило управляемость системы, но и предоставило ценный образовательный инструмент для демонстрации принципов наблюдаемости современных ИТ-систем.

Формирование культуры Infrastructure as Code в рамках образовательного учреждения привело к более систематическому и документированному подходу к управлению инфраструктурой. Все конфигурации хранятся в виде кода, что обеспечивает версионирование, возможность аудита изменений и воспроизводимость настроек.

Развитие компетенций административной команды в области современных технологий виртуализации, автоматизации и мониторинга. В ходе проекта сформирована команда специалистов, способных не только поддерживать существующую инфраструктуру, но и развивать её в соответствии с эволюцией технологий.

Несмотря на общую успешность проекта, в процессе реализации и эксплуатации инфраструктуры были выявлены определенные ограничения, которые необходимо учитывать при планировании дальнейшего развития системы.

Ограничения масштабируемости на уровне физических серверов являются наиболее существенным техническим ограничением. Текущая архитектура основана на одном физическом сервере, что создает естественный предел масштабирования по вертикали. При дальнейшем росте нагрузки потребуется расширение инфраструктуры с добавлением новых физических серверов и созданием кластера Proxmox.

Отсутствие автоматического масштабирования ресурсов требует ручного вмешательства администраторов для изменения конфигурации виртуальных машин при изменении нагрузки. Это ограничивает возможности оперативного реагирования на краткосрочные пики нагрузки и может приводить к временному снижению производительности системы.

Ограниченные возможности интеграции с облачными сервисами затрудняют создание гибридной инфраструктуры, которая могла бы использовать преимущества как локальных, так и облачных ресурсов. Текущая архитектура ориентирована на on-premise развертывание и не предусматривает прозрачной миграции сервисов между локальной и облачной средой.

Необходимость ручной настройки виртуальных машин для оптимальной работы с конкретными приложениями требует специфических знаний и увеличивает время, необходимое для развертывания новых сервисов. Полная автоматизация процесса оптимизации с учетом особенностей различных типов приложений не реализована.

Ограничения системы резервного копирования связаны с использованием встроенных механизмов Proxmox, которые не обеспечивают полной гибкости в управлении политиками хранения и репликации резервных копий. Для критически важных данных может потребоваться внедрение специализированных решений для резервного копирования.

Зависимость от квалифицированного персонала для управления инфраструктурой сохраняется, несмотря на высокий уровень автоматизации. Многие задачи, особенно связанные с оптимизацией производительности и диагностикой сложных проблем, требуют глубоких технических знаний и опыта работы с соответствующими технологиями.

Ограничения в безопасности связаны с отсутствием комплексной системы управления идентификацией и доступом, которая интегрировала бы все компоненты инфраструктуры. Каждый сервис использует собственную систему аутентификации, что усложняет управление доступом и увеличивает риск человеческих ошибок.

IV.III Система мониторинга на базе платформенных средств. Мониторинг микросервисной архитектуры. Интеграция систем мониторинга с системами машинного обучения

Современная парадигма разработки программного обеспечения всё чаще опирается на микросервисную архитектуру, что предъявляет особые требования к системам мониторинга. Разработка, развертывание и эксплуатация распределённых микросервисов, взаимодействующих друг с другом в режиме реального времени, требует комплексного, масштабируемого и адаптивного подхода к

наблюдаемости. Универсальная система мониторинга, представленная в данной работе, демонстрирует полное соответствие этим требованиям и эффективно решает вызовы, характерные для микросервисной среды. Одним из ключевых достоинств системы является полный охват всех уровней микросервисной архитектуры. На инфраструктурном уровне осуществляется мониторинг физических серверов, виртуальных машин и контейнеров, включая параметры загрузки, доступности и ресурсного потребления. Уровень оркестрации (например, Kubernetes) анализируется через специализированные экспортёры, позволяющие отслеживать состояние подов, деплойментов, namespace и других сущностей. На уровне самих микросервисов реализован сбор как технических, так и прикладных метрик, а на уровне бизнес-функций — агрегация данных, отражающих состояние и эффективность выполнения комплексных пользовательских операций, охватывающих несколько сервисов. Для визуального анализа и управления данными созданы специализированные дашборды, адаптированные под особенности микросервисной архитектуры. Они позволяют формировать как обзорное представление о состоянии всей системы, так и проводить детализацию по отдельным сервисам. Важной особенностью является построение графов зависимостей между микросервисами, что даёт возможность быстро выявлять цепочки вызовов, зоны деградации и потенциальные точки отказа. Отображение производительности взаимодействия между компонентами (время отклика, частота ошибок, нагрузка) существенно облегчает анализ проблем, связанных с сетевыми задержками и перегрузками. 85 Система поддерживает автоматическое обнаружение микросервисов благодаря динамическому сервис-дискавери на основе Kubernetes API. Это особенно важно в условиях частых релизов и масштабирования. При этом сбор метрик стандартизирован в соответствии с подходами типа RED (Rate, Errors, Duration), что позволяет унифицировать наблюдаемость, независимо от особенностей реализации конкретного сервиса. Метрики на уровне контейнеров включают использование CPU, памяти, дискового ввода-вывода, сетевых интерфейсов. Также реализован сбор информации о состоянии оркестратора — наличии ошибок, нарушениях SLA и отклонениях от ожидаемого состояния. Особое внимание уделено трассировке запросов, проходящих через цепочки микросервисов. Используемый механизм распределённого трейсинга позволяет отслеживать полное прохождение пользовательского запроса: от точки входа до конечной точки обработки. Это даёт возможность выявлять узкие места в архитектуре, замедляющие выполнение бизнес-операций. Дополнительно трассировки могут быть коррелированы с логами и метриками, что позволяет не просто увидеть, что запрос затормозил, но и понять, где именно и по какой причине возникла задержка или ошибка. был Для демонстрации возможностей системы в условиях приближённых к реальному бизнесу реализован практический кейс мониторинга микросервисного приложения электронной коммерции, состоящего из 15 сервисов, включая сервисы корзины, каталога, оплаты, аутентификации и управления заказами. Были настроены JMX экспортёры для сбора производственных метрик, внедрены трассировки для ключевых операций (например, добавление товара в корзину, оформление и оплата заказа), а также реализована агрегация на уровне бизнес-процессов — таких как общее время оформления заказа и процент успешных транзакций. Для разных категорий пользователей (разработчики, операторы, бизнес-аналитики) были созданы дашборды с фокусом на соответствующие метрики и параметры. Система алертинга была настроена на обнаружение деградации производительности отдельных сервисов и бизнес-операций, с дифференцированной маршрутизацией уведомлений. Внедрение системы в рамках кейса позволило достичь существенных операционных улучшений. Среднее время обнаружения инцидентов было сокращено на 67%, за счёт визуальной корреляции алертов, логов и трассировок. Время диагностики проблем сократилось на 58%, благодаря автоматической детализации причины сбоев. В результате была зафиксирована 43-процентная стабилизация системы: 86 снизилось число нештатных ситуаций, повысилась устойчивость микросервисного взаимодействия, ускорились циклы восстановления. Таким образом, представленная система мониторинга обеспечивает все критически важные элементы наблюдаемости в микросервисной архитектуре, сочетая техническую глубину, гибкость и адаптивность с высоким уровнем визуализации и аналитики, необходимыми для устойчивой и эффективной работы современных распределённых приложений.

Применение методов машинного обучения (ML) в универсальной системе мониторинга представляет собой важное направление её стратегического развития, позволяющее выйти за рамки реактивного наблюдения и перейти к проактивной, предиктивной и интеллектуальной модели мониторинга. Интеграция ML-компонентов расширяет функциональность платформы, повышает её адаптивность к сложным сценариям и обеспечивает новые возможности в сфере анализа и принятия решений. К числу перспективных областей применения машинного обучения в системе мониторинга относятся, прежде всего,

обнаружение аномалий в метриках и логах. Традиционные методы на основе жёстких порогов или статических правил зачастую недостаточно чувствительны к сложным или нестандартным отклонениям. ML-модели способны выявлять паттерны поведения, выходящие за рамки нормы, даже при отсутствии явно выраженных нарушений. Это особенно важно для многомерных метрик и логов с высоким уровнем шума. Не менее значимой задачей является прогнозирование поведения системы, включая предсказание роста нагрузки, вероятности отказов, исчерпания ресурсов и других событий. Использование моделей временных рядов, таких как LSTM или Prophet, позволяет создавать точные прогнозы на основе исторических данных, что критично для управления SLA, планирования масштабирования и оптимизации затрат. Дополнительно ML может применяться для автоматической классификации инцидентов — по типу, источнику, вероятной причине и приоритету. Это позволяет ускорить диагностику и маршрутизацию инцидентов к соответствующим командам. Агрегация инцидентов и событий на основе кластеризации или семантического анализа (например, с использованием NLP) помогает в борьбе с “шумными” алертами и повышает качество управления инцидентами. Интересное направление — рекомендательные модели, предлагающие оптимизации инфраструктуры, основанные на исторических шаблонах: например, перераспределение нагрузки, настройка порогов, масштабирование компонентов или 89 изменение конфигурации окружения. Также машинное обучение может быть использовано для выявления взаимосвязей между различными метриками, событийными потоками и бизнес-показателями, что даёт возможность более глубокой аналитики и построения причинно-следственных цепочек. Для реализации этой функциональности была спроектирована архитектура интеграции с ML-компонентами, включающая платформу для обучения и развертывания моделей (например, TensorFlow Serving или MLflow). Данные, поступающие в систему мониторинга (метрики, логи, трассировки), подвергаются сбору и подготовке — очистке, нормализации, сегментации по временным окнам и векторизации. После обучения модели развёртываются в режиме онлайн-инференса, с возможностью взаимодействия с компонентами мониторинга через RESTful или gRPC API. В архитектуре предусмотрен механизм обратной связи, позволяющий корректировать поведение моделей на основе реакций пользователей (например, подтверждение или отклонение алертов) и новых поступающих данных. Это обеспечивает непрерывное улучшение моделей, повышение точности и снижение количества ложных срабатываний. Среди конкретных моделей, рекомендованных для внедрения: Автоэнкодеры и изолирующие леса (Isolation Forest) — для детекции аномалий в высокоразмерных метриках; LSTM и Prophet — для прогнозирования трендов и нагрузки; NLP-модели (например, BERT) — для классификации и тематического анализа логов; K-means или DBSCAN — для кластеризации событий и автоматического объединения схожих инцидентов. Процесс внедрения ML-компонентов организуется поэтапно. Сначала производится обучение моделей на исторических данных из ClickHouse или других хранилищ. На следующем этапе осуществляется пилотное внедрение, в ходе которого ML работает параллельно с традиционными механизмами (например, пороговым алертингом), что позволяет сравнить результаты и оценить надёжность. Постепенно, по мере повышения доверия к модели и улучшения точности, она интегрируется в производственный пайплайн. Далее обеспечивается непрерывное обучение и обновление на новых данных, включая реализацию процессов retraining и CI/CD моделей. Таким образом, использование машинного обучения трансформирует систему мониторинга из инструмента наблюдения в интеллектуальную аналитическую 90 платформу, способную предсказывать, объяснять и предотвращать инциденты, адаптируясь к динамике как технических, так и бизнес-процессов. 9.2.2 Расширение функциональности Дальнейшее развитие универсальной системы мониторинга ориентировано на трансформацию её в комплексную платформу наблюдаемости, объединяющую технический и бизнес-контекст, способную обеспечить не только диагностику, но и предиктивный анализ и автоматическое реагирование. Основные направления эволюции платформы охватывают расширение функциональности в области трассировки, визуализации, интеграций, пользовательского взаимодействия и бизнес-аналитики. Одним из приоритетов становится усиление распределённой трассировки, что особенно актуально в условиях микросервисных и событийно-ориентированных архитектур. Планируется интеграция с OpenTelemetry в качестве стандартизированного фреймворка для сбора трассировок, метрик и логов. Это обеспечит унификацию данных и повысит совместимость с другими инструментами. Визуализация графов зависимостей между сервисами позволит наглядно представить взаимодействия между компонентами системы, а автоматический анализ трассировок — выявлять узкие места, высокую латентность и нестабильные точки без вмешательства пользователя. Развитие возможностей визуализации направлено на создание более интерактивного и адаптивного

пользовательского интерфейса. В рамках этих задач планируется внедрение интерактивных дашбордов с возможностью глубокой детализации, реализация механизмов визуализации топологии инфраструктуры и сервисов, а также генерация специализированных представлений для различных ролей — от операторов до бизнес-аналитиков. Отдельное внимание будет уделено автоматической генерации дашбордов на основе метаданных и конфигураций, что позволит сократить затраты на ручную настройку и повысить скорость адаптации системы под новые сервисы. В области интеграций планируется расширение взаимодействия с внешними системами. В частности, интеграция с системами управления инцидентами (например, Jira, Opsgenie, ServiceNow) позволит автоматически создавать и отслеживать тикеты по событиям мониторинга. Взаимодействие с системами управления конфигурациями(Ansible, Puppet, Terraform) обеспечит сквозной контроль от инфраструктуры до мониторинга. Интеграция с CI/CD пайплайнами даст возможность автоматически запускать тесты на мониторинг и метрики при каждом релизе. Также планируется синхронизация с системами автоскейлинга, что позволит адаптировать 91 масштабируемость приложений на основе реальных данных о нагрузке и производительности. Пользовательский опыт будет существенно улучшен за счёт разработки интуитивного интерфейса для настройки мониторинга, автоматизации рутинных операций (например, добавления новых целей, настройки алертов), а также внедрения интеллектуального поиска по логам, трассировкам и метрикам. Персонализированные представления, профили уведомлений и контекстные рекомендации повысят удобство и точность работы с системой для каждого пользователя. Особое внимание уделяется расширению возможностей бизнес-мониторинга. Разработка отраслевых метрикпозволит адаптировать систему под конкретные задачи — будь то финансовые транзакции, логистика, телеком или электронная коммерция. Интеграция с бизнес-системами (CRM, ERP) создаст единое аналитическое пространство, связывающее технические события с бизнес-процессами. Будет усилен блок отчетности и аналитики для бизнес-пользователей, включая визуальные отчёты, дашборды KPI и предиктивную аналитику. Ключевым направлением, объединяющим все уровни развития, станет внедрение методов машинного автоматического обнаружения обучения. Интеллектуальные аномалий, кластеризации модели для событий, предсказания отказов и рекомендаций по оптимизации инфраструктуры позволят перейти от реактивного мониторинга к проактивному управлению состоянием систем. Обратная связь от пользователей и контекстная адаптация моделей обеспечат непрерывное улучшение качества предсказаний. Таким образом, развитие универсальной системы мониторинга будет направлено на создание единой, сквозной платформы наблюдаемости, охватывающей технические, прикладные и бизнес-уровни. Такая система обеспечит глубокое понимание состояния всех компонентов ИТ-инфраструктуры, их взаимосвязей и влияния на ключевые бизнес-показатели, повышая не только технологическую устойчивость, но и управляемость цифровых процессов в организации.

Заключение и выводы

В работе представлена концепция и опытная эксплуатация «Цифрового полигона» как кампусной платформы, сочетающей виртуализацию, инфраструктуру как код и наблюдаемость для поддержки учебных, исследовательских и прикладных индустриальных сценариев. Продемонстрировано, что согласование академической и промышленной логик (через призму TRL) возможно при наличии прозрачных SLO/SLA, унифицированной телеметрии и строго регламентированных процессов развертывания и сопровождения. Можно зафиксировать, что:

1. «Цифровой полигон» выступает рабочей моделью «моста» между TRL 1–5 (университет) и TRL 6–9 (индустрия), снижая барьеры передачи результатов и ускоряя пилотирование сервисов.
2. Связка виртуализации (Proxmox VE) + IaC (Ansible) + наблюдаемость (Prometheus/Grafana, распределённый трейсинг) обеспечивает воспроизводимые деплои, управляемую конфигурацию и оперативную диагностику инцидентов.
3. Зафиксированы измеримые эксплуатационные эффекты: рост доступности до ~99,5%, сокращение инцидентов примерно на 78%, уменьшение TTD/TTR на микросервисном стенде на ~67% и ~58% соответственно, время развертывания типовых служб — 1–2 часа.
4. Унифицированная телеметрия с бизнес-метриками повышает наблюдаемость на уровне продукта (а не только инфраструктуры) и улучшает взаимодействие между разработкой, эксплуатацией и учебными группами.
5. Локальные сценарии ИИ (on-prem LLM на рабочих станциях с несколькими GPU) жизнеспособны для образовательных и пилотных задач и расширяют диапазон «безоблачных» применений.
6. Выявленные ограничения (одиночный сервер без кластеризации/HA, разрозненный IAM/SSO, базовые политики резервного копирования, отсутствие облачного бурста) не

носят принципиального характера и могут быть системно сняты.

7. Подход обладает высокой переносимостью: минимальные требования к оборудованию, открытые инструменты и регламенты позволяют реплицировать решение в других вузах и исследовательских центрах.

Перспективы развития Цифрового полигона ВИШ НИЯУ МИФИ:

- кластеризация и отказоустойчивость (HA) платформы, унификация IAM/SSO (например, Keycloak), политика бэкапов по схеме 3-2-1 с регулярной проверкой восстановления;
- формализация SLO/SLA по каждому сервису, автоматизация runbook'ов инцидентов, внедрение capacity planning;
- расширение контура наблюдаемости на бизнес-события и трассировку пользовательских сценариев (end-to-end);
- подготовка гибридной схемы с облачным бурстом для пиковых нагрузок и экспериментальных задач ИИ.
- предиктивная аналитика на телеметрии (прогнозные модели, детекция аномалий, приоритизация алертов);
- контролируемые эксперименты для факторного разложения вклада отдельных практик (виртуализация, IaC, типы трассировки) в итоговые KPI;
- публикация воспроизводимого набора артефактов (скрипты деплоя, конфигурации, тестовые нагрузки) для верификации результатов сообществом.

БИБЛИОГРАФИЯ

[1] AI Index Steering Committee. Artificial Intelligence Index Report 2025. Stanford University, Institute for Human-Centered AI (HAI), 2025. Доступно по ссылке: https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf. Дата обращения: 14.08.2025. Stanford HAI

[2] European Commission. Horizon Europe Work Programme 2021–2022: 13. General Annexes. Technology Readiness Levels (TRL). Luxembourg: Publications Office of the EU, 2022. Доступно по ссылке: https://ec.europa.eu/... (PDF). Дата обращения: 14.08.2025. European Commission

[3] Приказ Госкорпорации «Росатом» «Об утверждении перечня уровней готовности технологий и производства» от 24.04.2018 № 1/420-П (Приложение в редакции Приказа Госкорпорации «Росатом» от 11.08.2021 № 1/1007-П).

[4] National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST, 2023. DOI: 10.6028/NIST.AI.100-1. Доступно по ссылке: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. Дата обращения: 14.08.2025. NIST Publications

[5] Google SRE Team. Service Level Objectives (глава онлайн из «Site Reliability Engineering: How Google Runs Production Systems»). O'Reilly Media, 2016 (онлайн-версия на sre.google). Доступно по ссылке: https://sre.google/sre-book/service-level-objectives/. Дата обращения: 14.08.2025. sre.google

[6] OpenTelemetry Project (CNCF). OpenTelemetry Specification, v1.47.0 (Overview). 2025. Доступно по ссылке: https://opentelemetry.io/docs/specs/otel/. Дата обращения: 14.08.2025. OpenTelemetry

[7] Wilkie, T. The RED Method: How to Instrument Your Services. Grafana Labs Blog, 02.08.2018. Доступно по ссылке: https://grafana.com/blog/2018/08/02/the-red-method-how-to-instrument-your-services/. Дата обращения: 14.08.2025. Grafana Labs

[8] Zhou, Z.; Ning, X.; Hong, K.; et al. A Survey on Efficient Inference for Large Language Models. arXiv:2404.14294, 2024. Доступно по ссылке: https://arxiv.org/abs/2404.14294 (PDF: https://arxiv.org/pdf/2404.14294). Дата обращения: 14.08.2025.

Статья получена 6 августа 2025.

Гаджилов Гамзат Гаджиевич, выпускник магистратуры ВИШ НИЯУ МИФИ, gadjilov@bk.ru

Хатунов Амгалан Владимирович, выпускник магистратуры ВИШ НИЯУ МИФИ, fanqo@yandex.ru

Волошин Тарас Андреевич, аспирант ВИШ НИЯУ МИФИ, tvoloshin38@gmail.com

Жабицкий Михаил Георгиевич, заместитель директора ВИШ НИЯУ МИФИ, jabitsky@mail.ru

MEPhI Higher Engineering School Digital polygon for educational and practical projects infrastructure support

G.G. Hajilov, A.V. Khatunov, T.A. Voloshin, M.G. Zhabitsky

Abstract: ***The article describes the concept, architecture and results of the trial operation of the "Digital Polygon" of the Higher School of National Research Nuclear University MEPhI - a digital infrastructure for combining education, research and applied projects with industrial partners. It is shown how the gap in expectations between academic and industrial environments (in terms of levels of technological readiness, TRL) is transformed into requirements for services, access policy, and observability. A target architecture based on virtualization (Proxmox VE), infrastructure as code (Ansible) and a single observability loop (Prometheus/Grafana, distributed tracing) is proposed, as well as operating regulations for educational research and pilot production scenarios. The cases demonstrated: local inference of large language models on workstations with multiple GPUs and a microservice application bench (≈15 services) with tracing and business metrics. As a result of the implementation, improvements in operational indicators were recorded: an increase in availability up to ~99.5%, a decrease in the number of incidents by ~78%, a reduction in the deployment time of typical services to 1-2 hours; for a microservice bench – a reduction in the time of detection and elimination of failures (TTD/TTR) by 67% and 58%, respectively. The scientific and practical novelty of the work lies in the integration of the TRL approach to interaction with the industry with a reproducible engineering template of the campus platform (virtualization + IaC + observability) and the demonstration of its applicability for on-prem AI tasks. Further development steps are outlined: clustering, IAM/SSO unification and backup policy.***

The keywords are digital polygon, virtualization, monitoring, infrastructure as code, observability, microservice architecture, scalability, fault tolerance.

REFERENCES

[1] AI Index Steering Committee. Artificial Intelligence Index Report 2025. Stanford University, Institute for Human-Centered AI (HAI), 2025. Доступно по ссылке: https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf. Stanford HAI

[2] European Commission. Horizon Europe Work Programme 2021–2022: 13. General Annexes. Technology Readiness Levels (TRL). Luxembourg: Publications Office of the EU, 2022. // https://ec.europa.eu/... (PDF). Дата обращения: 14.08.2025. European Commission

[3] Order of Rosatom State Corporation "On Approval of the List of Levels of Readiness of Technologies and Production" dated 24.04.2018 No 1/420-P (Appendix as amended by the Order of Rosatom State Corporation dated 11.08.2021 No 1/1007-P).

[4] National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST, 2023. DOI: 10.6028/NIST.AI.100-1. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

[5] Google SRE Team. Service Level Objectives (глава онлайн из «Site Reliability Engineering: How Google Runs Production Systems»). O'Reilly Media, 2016 https://sre.google/sre-book/service-level-objectives/.

[6] OpenTelemetry Project (CNCF). OpenTelemetry Specification, v1.47.0 (Overview). 2025. Доступно по ссылке: https://opentelemetry.io/docs/specs/otel/.

[7] Wilkie, T. The RED Method: How to Instrument Your Services. Grafana Labs Blog, 02.08.2018. Доступно по ссылке: https://grafana.com/blog/2018/08/02/the-red-method-how-to-instrument-your-services/.

[8] Zhou, Z.; Ning, X.; Hong, K.; et al. A Survey on Efficient Inference for Large Language Models. arXiv:2404.14294, 2024.