



Обзор конференции "Computing in High Energy and Nuclear Physics" - CHEP-2026

<https://indico.cern.ch/event/1471803/>

Reporter: Andrey Shevel

Plan

- Информация о конференции
- Основные новые темы (почти везде на основе ИИ)
 - Автономные агенты
 - Выводы на основе симуляции
 - (Simulation-Based Inference)
 - Сквозной дифференцируемый конвейер
 - (End-to-End differentiable pipeline)
 - Прогнозы по объёмам данных
 - Прогнозы запросов на вычисления
 - Проект для анализа данных
 - (Dedicated Analysis Facilities (Afs))
- Разработанный цифровой двойник (ЦД)

История конференции

- Серия конференций CHEP стартовала в 1985 году.
- Проводится примерно раз в 18 месяцев в разных странах.
- Обычное число участников свыше 500.
- CHEP-2026 в Бангкок/Таиланд (25-29 мая) <https://indico.cern.ch/event/1471803/>

Тематика: использование компьютерной инфраструктуры в крупных экспериментах

- Track 1 – Data and Metadata Organization, Management, and Access (50)
- Track 2 – Online and Real-time Computing (55)
- Track 3 – Offline Data Processing (54)
- Track 4 – Distributed Computing (44)
- Track 5 – Event Generation and Simulation (42)
- Track 6 – Software Environment and Maintainability (33)
- Track 7 – Computing Infrastructure and Sustainability (60)
- Track 8 – Analysis Infrastructure, Outreach, and Education (30)
- Track 9 – Analysis Software and Workflows (44)

Oral выступления из РФ

- Дубна (ОИЯИ) – 2
- Дубна + Москва (ВШЭ) – 1
- Томск – 1
- Гатчина – 1

Автономный ИИ Агент

- ИИ Агент — это программная система, которая использует ИИ (обычно в форме LLM) для автономного достижения целей и выполнения задач от имени пользователей, обладая способностью к рассуждению, планированию, запоминанию, а также принятию решений, обучению и адаптации.

Ключевые фреймворки автономных агентов

- **Dr.Sai:** новаторская многоагентная система в BESIII, которая обрабатывает такие задачи, как отбор и подгонка данных. Она предоставляет план сотрудничества человека и ИИ, где агент выступает в качестве активного партнера в научном мышлении. - <https://indico.cern.ch/event/1471803/contributions/6967082/>
- **LHCb Brain:** агентный помощник, разработанный для поддержки операций, разработки программного обеспечения и анализа. Он использует интеллектуальный маршрутизатор на основе LLM для направления запросов к специализированным серверам протокола контекста модели (MCP) и внутренним службам для выполнения задач, а не просто генерации текста. <https://indico.cern.ch/event/1471803/contributions/6966788/>
- **Archi:** фреймворк с открытым исходным кодом для создания агентов ИИ, используемый группой вычислительных операций CMS. Он автоматизирует исследовательские и операционные рабочие процессы, объединяя документацию, код и диагностику в реальном времени, чтобы помочь операторам и снизить нагрузку на экспертов. <https://indico.cern.ch/event/1471803/contributions/6967119/>
- **SciBot:** специализированный помощник для программы сохранения данных и анализа RHIC. Он использует архитектуру Retrieval-Augmented Generation (RAG) для обеспечения безопасного доступа на естественном языке к десятилетиям сохраненных научных знаний. <https://indico.cern.ch/event/1471803/contributions/6966830/>

Examples

- **Workflow Management:** Researchers are exploring "agentic AI" interfaces for distributed High Throughput Computing. These agents can leverage the Model Context Protocol (MCP) to submit jobs, check status, and transfer data, closing the gap between complex infrastructure and scientific output
- **Operations and Diagnostics:** At Belle II, AI agents are being explored for automated analysis of grid job logs, failure classification, and root-cause suggestion to reduce service downtime
- **Metadata Automation:** The Metadex project for DUNE uses a multi-agent architecture to automate metadata pipelines, including workflow planning, flagging anomalies, and dynamic query construction from natural language
- **Collaborative Interaction:** A new paradigm is shifting away from traditional GUI-centric workflows toward a dialogic interaction where users collaborate with software via conversation, creating a feedback loop of "what you say is what you see"

Underlying Technologies

- These agents typically rely on Large Language Models (LLMs) integrated with experiment-specific knowledge bases via Retrieval-Augmented Generation (RAG).
- While these agents act with high degrees of autonomy, the community emphasizes the continued necessity of human-in-the-loop controls for security-critical workflows.

Simulation-Based Inference

- Simulation-Based Inference (SBI), and specifically Neural Simulation-Based Inference (NSBI), is emerging at CHEP 2026 as a critical methodology for statistical interpretation.

https://indico.cern.ch/event/1471803/contributions/6968234/attachments/3283447/5868309/SBI_Toolkit_CHEP.pdf

- Simulation-Based Inference (SBI) — also called likelihood-free inference — is the family of methods you use when you have a stochastic simulator $x \sim p(x | \theta)$ (e.g. Pythia + Geant4 + reconstruction) but no tractable likelihood $p(x | \theta)$. Instead of evaluating the likelihood.

Key Frameworks and Toolkits

- **NEEDLE**: A columnar workflow orchestrator designed for large-scale NSBI . It uses a directed acyclic graph (DAG) workflow manager for training/evaluation and employs Dask-based libraries for efficient memory management when reading ROOT and Parquet formats. <https://indico.cern.ch/event/1471803/contributions/6968231/>
- **nsbi-common-utils**: An open-source Python toolkit that provides a modular, end-to-end workflow for NSBI, including data preparation, model training, and statistical inference. <https://indico.cern.ch/event/1471803/contributions/6968234/>
- **everwillow (JAX Ecosystem)**: A model-agnostic inference tool that leverages JAX for differentiable, likelihood-based inference. It allows analyses to benefit from hardware acceleration, analytical gradients, and live fitting visualization. <https://indico.cern.ch/event/1471803/contributions/6968265/>
- **Differentiable RooFit**: Researchers have demonstrated end-to-end SBI workflows within the RooFit framework by integrating neural surrogates. This approach uses Automatic Differentiation (AD) via tools like Clad to calculate exact gradients of the full likelihood with respect to model parameters. <https://indico.cern.ch/event/1471803/contributions/6968265/>

Representative Applications

- **Precision Flavour Physics (Belle II):** SBI is used to measure deviations from the Standard Model in semileptonic $B \rightarrow K^* \ell^+ \ell^-$ decays. Neural network approaches, including Deep Sets to enforce event-level permutation invariance, allow for meaningful parameter recovery even with detector smearing and background contamination.
- **Neutrino Physics (JUNO):** The JUNO experiment employs NSBI for Neural Monte Carlo Tuning. By using transformer-based and normalizing flow-based density estimators, they can approximate intractable likelihoods to recover energy response parameters with near-zero systematic bias.
- **Collider Physics (CMS):** CMS is leveraging NSBI for Effective Field Theory (EFT) analyses involving top pair production. This approach is particularly effective when many parameters of interest make hand-engineered variables insufficient for inferring the likelihood.

End-to-End differentiable pipeline

- An end-to-end differentiable pipeline in high-energy physics (HEP) is a modular analysis framework where every stage—from event selection and observable construction to histogramming and final likelihood-based inference—is mathematically differentiable.
- This enables the use of gradient-based optimization across the entire workflow, allowing analysts to systematically tune every parameter of an analysis simultaneously to maximize statistical sensitivity or minimize systematic uncertainties.

Hardware Optimization and Detector Design

- The importance of differentiability extends beyond data analysis to the physical design of experiments:
 - <https://indico.cern.ch/event/1471803/contributions/6968005/>
- **Differentiable Simulation:** Applying algorithmic differentiation to tools like Geant4 enables gradient-based optimization for detector hardware parameters and design.
- **Interoperability:** Most differentiable ecosystems are built on JAX, which provides multi-hardware acceleration (GPUs/TPUs), allowing these complex optimizations to run efficiently on modern heterogeneous computing resources.

Forecast for Data Volumes

- HL-LHC (Run-4 and Run-5): In Run-4 (starting ~2030), data transfer rates and computing resource needs are expected to grow by an order of magnitude (10x) compared to Run-3. By Run-5, these demands are anticipated to double again as the LHC reaches its ultimate luminosity goals.
- Exabyte-Scale Storage: Major experiments like ATLAS have already surpassed 1 exabyte of stored data. HL-LHC operations are expected to generate scientific data on the scale of multiple exabytes annually.
- DUNE: The Deep Underground Neutrino Experiment is projected to produce approximately 30 PB of raw data per year from just its first two Far Detector modules. Single data records for readouts will reach GB scales, orders of magnitude larger than current collider event sizes.
- Spin Physics Detector (SPD): This experiment at the NICA complex expects a raw data stream of 20 GB/s, resulting in an annual volume of 10 PB of preprocessed data.
- Einstein Telescope (ET): As a next-generation gravitational wave observatory, ET is expected to require computing and data resources on a scale similar to LHC experiments.

Forecast for Computing Power

- The "one-size-fits-all" approach to computing is being replaced by heterogeneous architectures to handle the massive increase in processing demands:
 - Scaling of Workloads: The number of jobs processed by accounting systems like AUDITOR is expected to increase by a factor of 7 to 10 for the HL-LHC era. Similarly, the number of concurrent jobs at CERN's Tier-0 is expected to rise by a factor of 4 to 7, growing from the current 125,000 to potentially nearly a million.
 - End of x86 Dominance: The HEP community is entering a "Resource Renaissance" defined by the proliferation of ARM, RISC-V, and diverse GPU generations.
 - Experiments are increasingly offloading reconstruction and simulation tasks to GPUs to achieve the necessary throughput and energy efficiency.
 - Real-Time Processing at 40 MHz: Next-generation triggers must perform offline-quality reconstruction in real-time at 40 MHz, necessitating massive use of FPGAs and AI Engines to manage sub-microsecond latency requirements.

Mitigation and Efficiency Strategies

- To manage these projections within feasible budget and energy constraints, the community is adopting several innovations:
 - Data Format Transition: Transitioning from ROOT TTree to the new RNTuple format is delivering a 50% reduction in storage footprint and faster I/O performance.
 - Networking Upgrades: Wide Area Networks (WAN) are being upgraded to 400 Gbps and beyond.
 - There is a concerted effort to remove IPv4 entirely from WLCG links by the end of 2026 to support the horizontal scaling of storage servers required for HL-LHC.
 - AI-Driven Reduction: Machine learning techniques, such as GANs and Diffusion Models, are being used for fast simulation to reduce the CPU hours spent on traditional Monte Carlo production by several orders of magnitude.
 - Energy-Aware Computing: Sites are implementing predictive models and CPU frequency scaling to modulate power draw in alignment with grid carbon intensity and electricity costs.

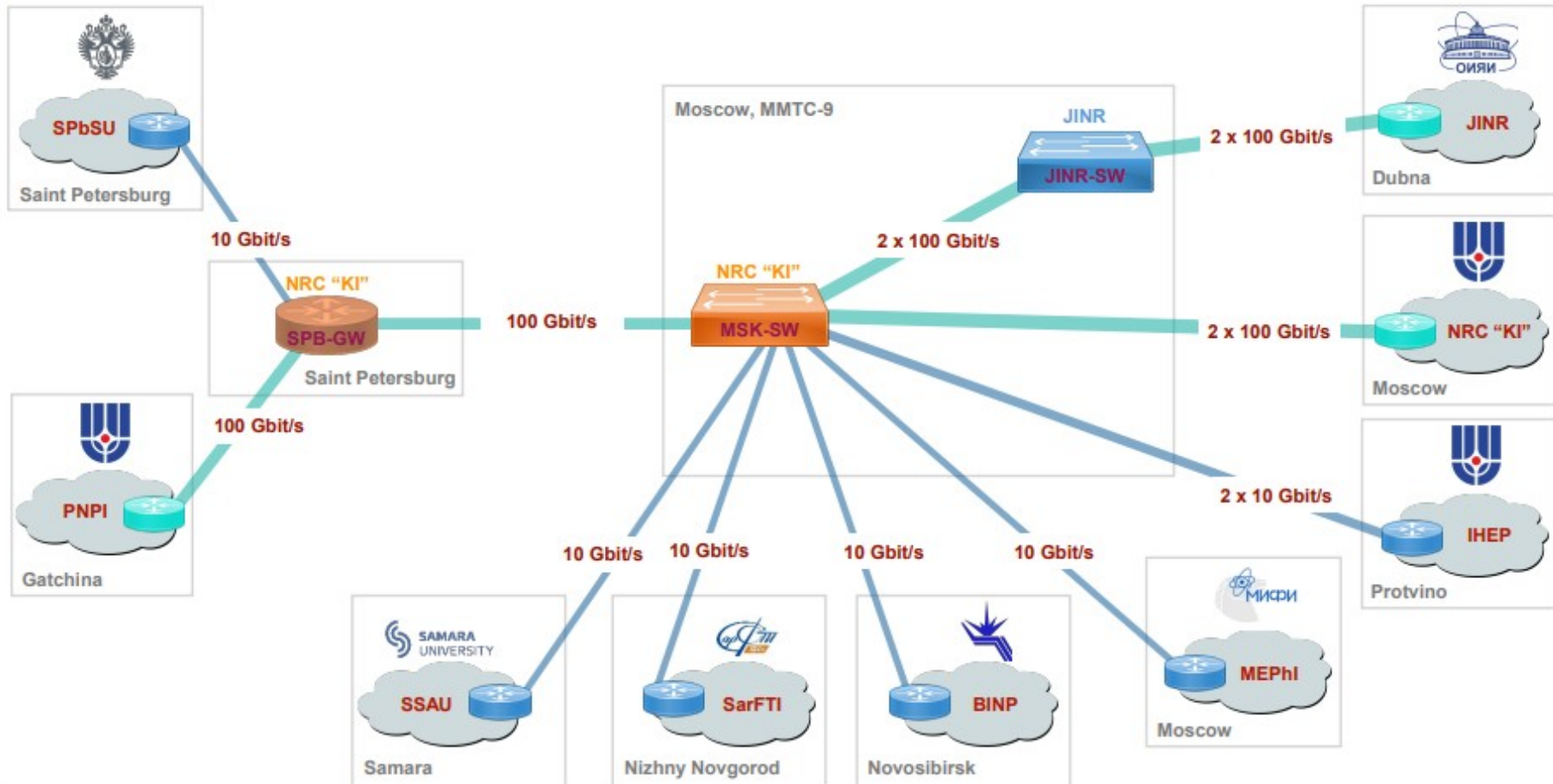
Dedicated Analysis Facilities (AFs)

- **CERN Analysis Facility:** Evolves SWAN (CERN's web-based analysis environment) to support columnar workloads using Dask to manage concurrent resources from the batch farm.
<https://indico.cern.ch/event/1471803/contributions/6968292/>
- **Purdue Analysis Facility:** A Kubernetes-based platform for CMS researchers that integrates with LLM-enabled IDEs and agentic tools to assist with code development and debugging.
<https://indico.cern.ch/event/1471803/contributions/6968305/>
- **Next Generation (Triggers) Platform:** A CERN project that unifies traditional workflows with a centralized pool of accelerator resources (GPUs/FPGAs). Users access these interactively through SSH, notebooks, and VSCode.
<https://indico.cern.ch/event/1471803/contributions/6966769/>

Integration and Orchestration Tools

- **interLink:** A technology that allows a workstation or local Kubernetes cluster to transparently offload payloads to remote supercomputers (HPC centers) while preserving a local-like interactive feel.
<https://indico.cern.ch/event/1471803/contributions/6966738/>
- **fapptainer:** Enables researchers to run unchanged experiment-specific software (distributed via CVMFS) on varied HPC resources from their interactive environment.
<https://indico.cern.ch/event/1471803/contributions/6966454/>

SPD computing network



Andrey Kiryanov, Danila Oleynik, Artem Petrosyan



The digital twin for server network analysis with LLM- driven RAG integration

Reporter: Andrey Shevel

Co-authors: Alexey Naikov, Anatoly Oreshkin, Alexey Shvetsov

<https://indico.cern.ch/event/1471803/contributions/6966847/>

The Digital Twin (DT)

- Natural-language interaction with DT.
- Specialized command set and extensions with custom options to query the live monitoring data.
 - These requirements motivate the use to of locally embedded large language model (LLM) features to query the ontology and operational data.

Как использовать LLM

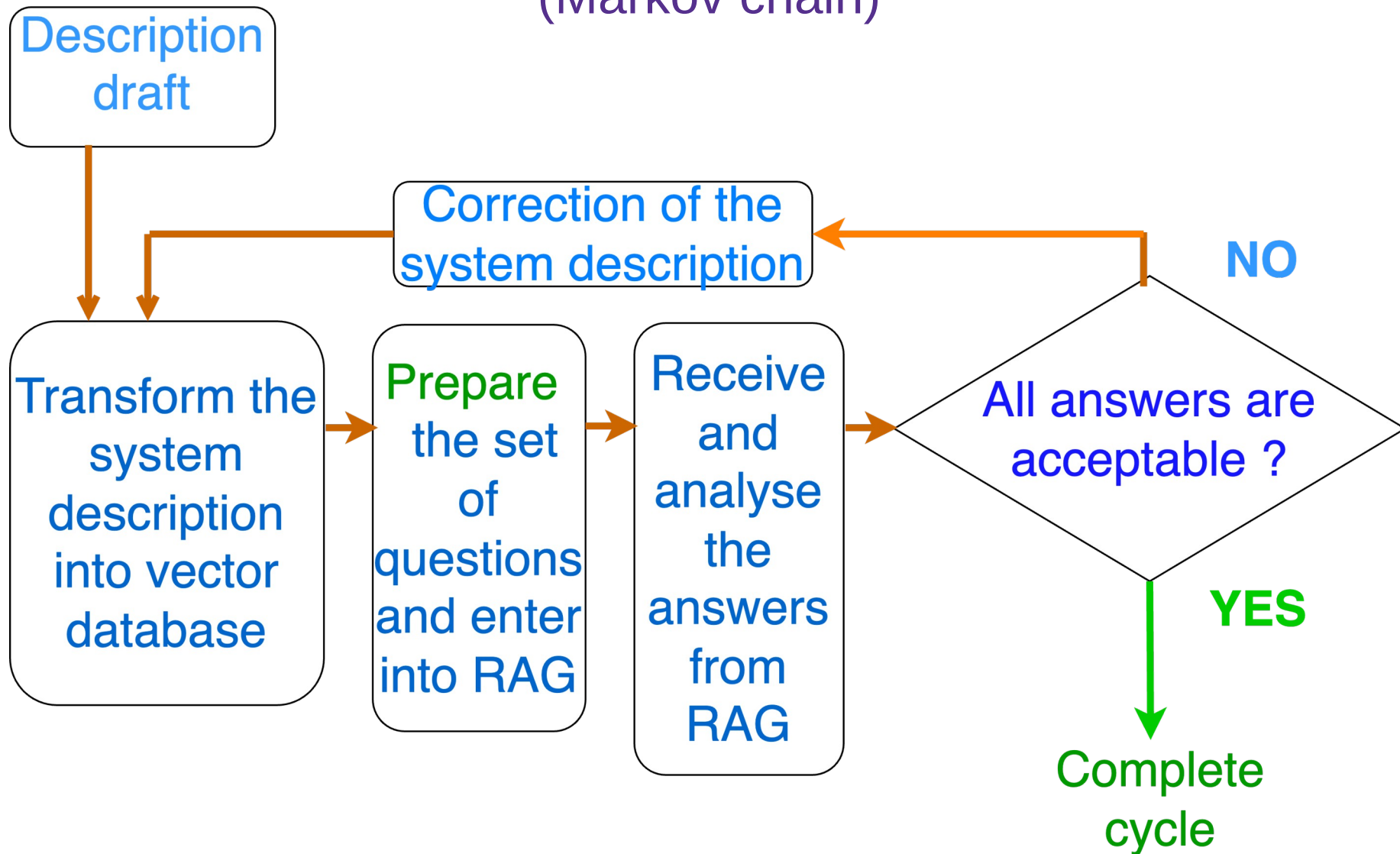
- Без строгого контроля LLMs склонны к галлюцинациям.
- Чтобы минимизировать вероятность галлюцинаций предприняты важные шаги:
 - Использован фреймворк “Поиск в онтологии и дополнительная генерация LLM” (Retrieval Augmented Generation - RAG).
 - Температура генерации LLM = 0.

Разработка онтологии

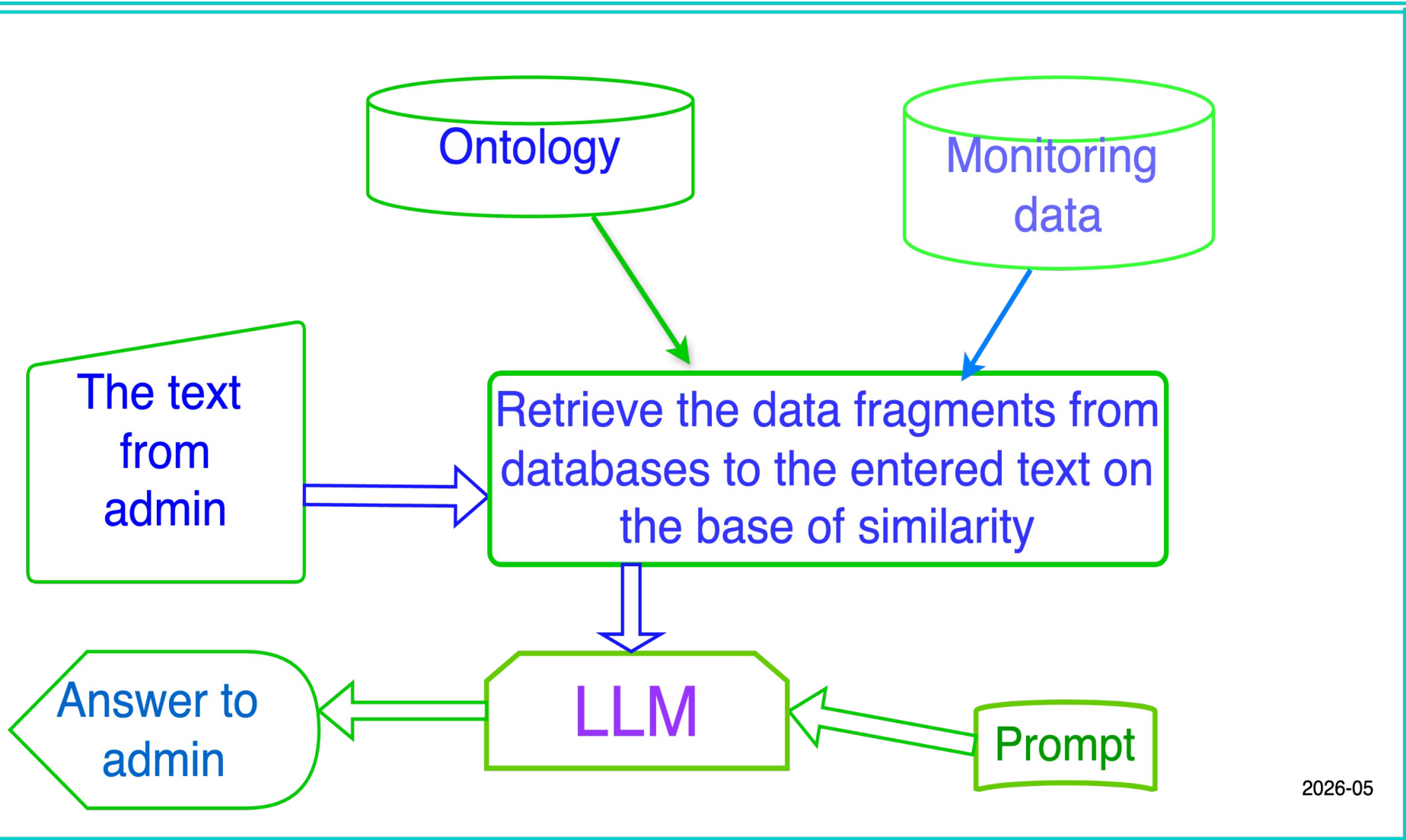
- Простейшая форма онтологии – это описание объекта на естественном языке.
 - Сама разработка онтологии выполнялась (показано на следующем слайде) посредством специального цикла с использованием фреймворка RAG.
- В дополнение использована более строгая форма онтологии, где система описана на языке OWL, с использованием приложения Protégé. [Enabling Operational Data Analytics for Datacenters through Ontologies, Monitoring, and Simulation-based Prediction <https://doi.org/10.1145/3629527.3652897>].

Cycle to develop/correct the text description

(Markov chain)

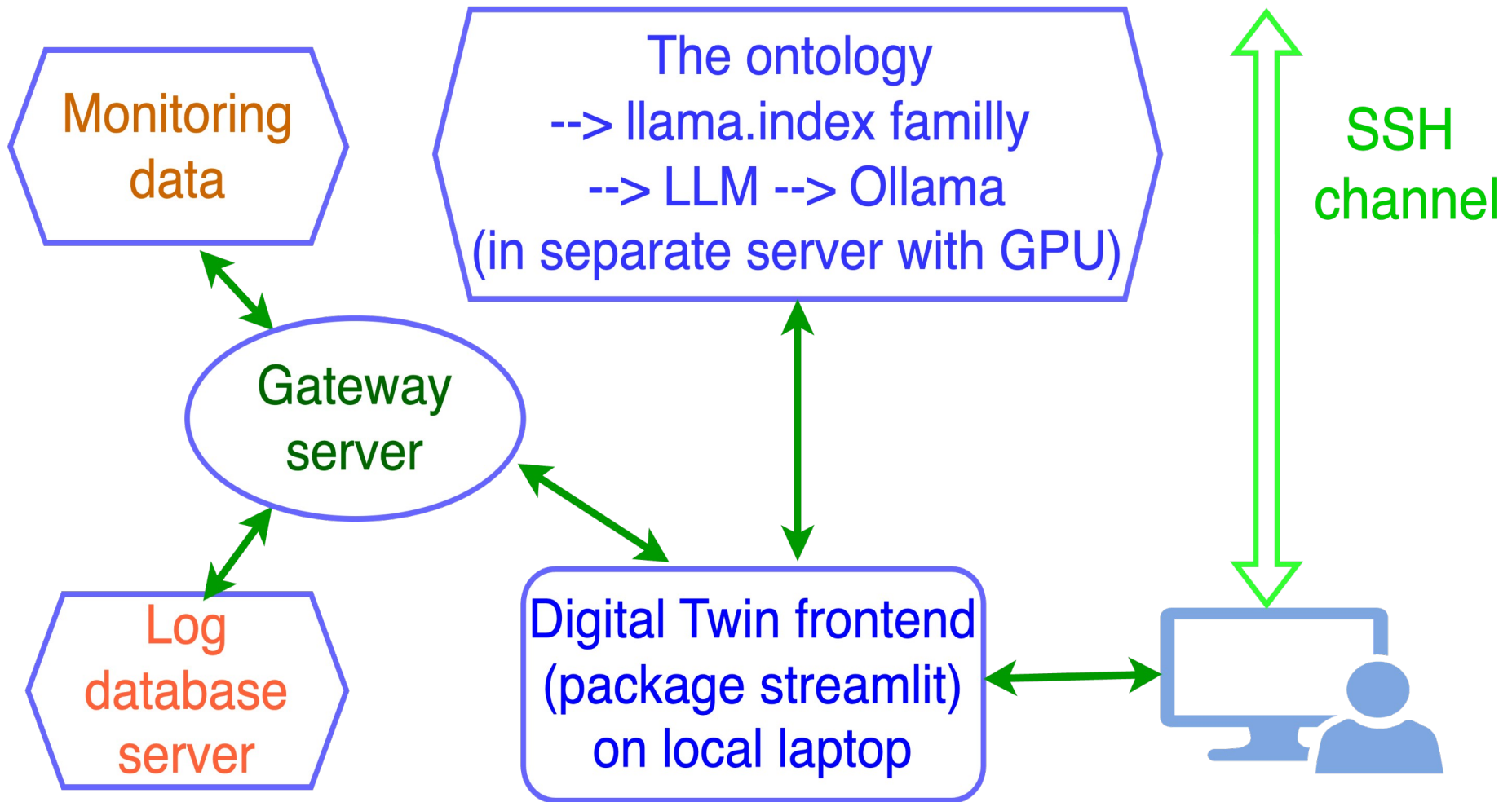


Digital twin logic



2026-05

Implemented digital twin architecture



2026-05

Предварительные результаты

- Реализованный прототип ЦД показал следующее:
 - “Быстрый” ответ (1+ минут) на подготовку ответа на введённый вопрос администратора (описание сети – около 200 страниц).
 - Показывает текущее состояние запрошенного сервера.
 - Возможность интерпретации сообщения из системного лога с использованием средств ЦД.
- В целом: ЦД помогает опытному администратору.

Example of the answer from DT

You entered: 2026-05-27 14:44:15 — what is dbserve?

Completion code: 200

Any answer of digital twin must be checked!

The answer took around 1.7 minutes or 100.0 seconds

Streamlit: The frontend has been performed on MacBook-Pro.local

2026-05-27 14:45:55 — The digital twin answer: {"answer":"Based on the provided documentation, **dbserve** (specifically `dbserve.pnpi.spb.ru`) is a virtual server hosted on the physical machine `new-dbserve.pnpi.spb.ru`. Key details from the text include:

- **Purpose:** Serves as the web server for ОФБЭ.
- **Operating System:** AlmaLinux release 9.4 (Seafoam Ocelot).
- **Network & Aliases:** Has the alias `hepd.pnpi.spb.ru` and a secondary network interface with IP `192.168.14.195` (DNS: `dns2.pnpi.spb.ru`).
- **Host Environment:** Automatically starts via KVM on the physical host `new-dbserve.pnpi.spb.ru`.
- **Installed Services:** Runs Apache 2.4.57, PostgreSQL 13.16, Postfix 3.5.9, BIND 9.16.23, Mailman 3.3.8,

Стоимость реализации ЦД

- Реализация ЦД требует трудозатрат от команды администраторов + оборудование (сервер, GPU(s)).
 - Имеем баланс:
 - Однажды разработать ЦД или нанять дополнительный штат на время жизни серверной сети.
- Однако ЦД может быть разработан как часть плановых работ по модернизации системы.

Трудозатраты по поддержке ЦД

- **Ручная работа:**
 - Коррекция онтологии при любых изменениях в сети.
 - Чем больше деталей в онтологии, тем точнее ответы. Но и больше трудозатрат на разработку онтологии.

Как оценить влияние ЦД

- Сравнить обычные процедуры администраторов с использованием и без использования ЦД.
- Опрос администраторов об их впечатлении от использования ЦД.

Уроки

- Цифровой двойник, который настроен на конкретный круг задач, с использованием LLM - отличный помощник администратора.
- Качество помощи цифрового двойника в значительной степени определяется качеством онтологии.
- Наличие GPU необходимой производительности является одним из предварительных условий.
- Процесс построения цифрового двойника может открыть способы улучшения самой системы — ЦД производит воздействие на оригинал.

Заключение

- Представлен работающий прототип цифрового двойника помощника администратора для управления серверной сети из 10 хостовых серверов, 25 виртуальных машин, которые выполняют десятки функций для ОФВЭ.
- Авторы демонстрируют подход, который может быть масштабирован для более сложных объектов:
 - Большие экспериментальные системы такие как
 - детекторы,
 - ускорители,
 - Worldwide LHC Grid (WLCG),
 - Future distributed physics ML systems.
 - Такой подход станет стандартом в обозримом будущем.

Questions?

Обзорная информация материалов СНЕР2026

- По ссылке <https://hepd.pnpi.spb.ru/CSD/CHEP-2026/> можно найти несколько обзорных файлов материалов СНЕР2026, которые я подготовил с помощью средств машинного обучения.

AI в сообщениях

- Machine learning – 96
- AI – 53
- Artificial Intelligence – 28
- LLM – 20
- Digital Twin - 4