И. В. КОТЕНКО, И. Б. САЕНКО
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук,
Санкт-Петербург

КОНЦЕПЦИЯ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ БОЛЬШИХ МАССИВОВ ГЕТЕРОГЕННЫХ ДАННЫХ О СОБЫТИЯХ КИБЕРБЕЗОПАСНОСТИ В КРИТИЧЕСКИ ВАЖНЫХ ИНФРАСТРУКТУРАХ

Рассмотрена концепция аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности, производимой в интересах оценки состояния, поддержки принятия решений и расследования компьютерных инцидентов в критически важных инфраструктурах. Проанализированы подходы к применению суперкомпьютерных вычислений для обработки больших данных по кибербезопасности. Предложена и обсуждена архитектура системы аналитической обработки гетерогенных данных о событиях кибербезопасности в критически важных инфраструктурах.

Введение. Рост объемов собираемых данных о событиях кибербезопасности и развитие средств и методов их надежного хранения привели к повышению актуальности разработки новых методов и алгоритмов анализа больших и сверхбольших наборов данных такого класса. Анализ таких объемов данных о событиях кибербезопасности в современных критически важных инфраструктурах (КВИ) требует привлечения технологий и средств реализации высокопроизводительных (суперкомпьютерных) вычислений [1]. Именно на этом пути необходима разработка с единых системных позиций концептуальных основ (концепции) аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности [2, 3].

Предлагаемый доклад посвящен рассмотрению концепции, которая призвана сформировать систему взглядов, идей и принципов аналитической (суперкомпьютерной) обработки больших данных в интересах оценки состояния, поддержки принятия решений и расследования компьютерных инцидентов в КВИ [4–6]. Она предназначена для заказчиков, разработчиков и пользователей перспективных систем мониторинга и управления безопасностью КВИ, в которых осуществляется аналитическая обработка больших массивов гетерогенных данных о событиях кибербезопасности в интересах оценки состояния, поддержки принятия решений и расследования компьютерных инцидентов в критически важных инфраструктурах. Концепция является методологической базой нормативно-технических и методических документов, направленных на решение следующих задач: выработку технических заданий на построение систем аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности в интересах оценки состояния; поддержку принятия решений и расследования компьютерных инцидентов в КВИ; создание систем аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности; сертификацию систем такого вида.

Подходы к применению суперкомпьютерных вычислений для обработки больших данных по кибербезопасности. Суперкомпьютерные вычисления являются важнейшей технологией для получения технологического лидерства в наукоемких разработках. Это касается самых разных областей деятельности человека, включая машиностроение, медицину, кибербезопасность, экономику, биоинформатику и многие другие. Одновременно с ростом объемов данных развиваются методы и технологии их анализа, в том числе методы извлечения новых знаний из массивов данных. Среди них можно выделить методы искусственного интеллекта, топологического анализа данных, статистического анализа, алгоритмы параллельной обработки данных, технологии потоковой обработки данных, и другие.

Особенностью суперкомпьютерных систем является их ориентация на параллельно-массивную обработку данных и решение масштабных задач [7]. Современные суперкомпьютерные системы построены по кластерной архитектуре, то есть состоят из большого количества вычислителей, мощных серверов с гомогенной или гетерогенной архитектурой. Под гомогенной архитектурой подразумевается наличие процессоров одного типа в одном вычислительном узле, например, Intel или ARM архитектуры. Под гетерогенной архитектурой подразумевается наличие процессоров нескольких типов в одном вычислительном узле, например Intel и GPU

одновременно. Все вычислители суперкомпьютерных систем объединены быстрой сетью, например, Infiniband. В отличие от сетей Ethernet, широко применяемых в дата-центрах и облачных системах, Infiniband или аналогичные технологии обеспечивают большую пропускную способность и низкую латентность (менее микросекунды) при обмене информации. Это дает больше возможностей для масштабирования при решении ресурсоемких вычислительных задач. Кроме того, применение гетерогенных вычислителей с использованием GPU или ПЛИС-ускорителей позволяет существенно повысить скорость обработки данных, а также обеспечить решение задач в области машинного обучения.

Задачи обработки больших данных по кибербезопасности можно отнести к следующим категориям: (1) обучение интеллектуальных алгоритмов; (2) масштабная обработка данных; (3) моделирование.

В рамках первой категории необходимо применение гетерогенных высокопроизводительных систем, оборудованных графическими ускорителями для обеспечения процессов обучения. Для решения задач глубокого машинного обучения и применения методов искусственного интеллекта эффективными инструментами являются графические ускорители с поддержкой технологии CUDA. Практически все современные библиотеки, например, TensorFlow, Keras, PyTorch и Theano, поддерживают данную библиотеку.

Ко второй категории относятся задачи обработки больших массивов данных с целью их предварительной подготовки для дальнейшего анализа, применения алгоритмов связывания данных, решения ресурсоемких задач по поиску сложных зависимостей в данных. Для эффективной обработки данных необходимо как можно большую их часть загрузить в память системы, чтобы повысить скорость их обработки. Для этой цели применяется кластерная архитектура суперкомпьютерных систем и используется технология MPI (Message Passing Interface). В результате можно построить программную среду с распределенной памятью, в которой обработчики (MPI-процессы) могут работать как со своими локальными данными, так и с данными других процессов, обмениваясь друг с другом сообщениями. Такой подход позволяет построить сложные структуры данных большого объема, в том числе в виде графов.

Третья категория относится к задачам моделирования. Спектр решаемых здесь задач - достаточно широкий, например задачи моделирования для проектирования комплексного применения отдельных компонентов защиты [8]. Отдельно можно выделить задачи имитационного моделирования, которые можно масштабировать [9]. Подход с использованием MPI позволяет применить акторную модель при моделировании и при этом поместить в каждый актор достаточно ресурсоемкий алгоритм. Кроме имитационного моделирования, актуально применение точных математических, статистических и других моделей для задач обработки, дополнения, восстановления и генерации данных.

Архитектура системы аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности (САО БМГД СБ). В разработанной архитектуре САО БМГД СБ (рисунок) выделяются следующие функциональные компоненты: (1) обнаружения в реальном времени атак на основе имитационного и графо-ориентированного моделирования; (2) обнаружения в реальном времени аномальной активности и нарушений критериев и политик безопасности; (3) оперативной оценки защищенности информационных, телекоммуникационных и других критически важных ресурсов; (4) оперативного анализа и управления рисками информационной безопасности; (5) оперативной визуализации больших массивов гетерогенных данных о событиях кибербезопасности; (6) принятия решений по защите информационных, телекоммуникационных и других критически важных ресурсов; (7) проведения расследований компьютерных инцидентов на основе аналитической обработки больших массивов гетерогенных данных о кибербезопасности; (8) взаимодействия с суперкомпьютерным вычислительным центром (СВЦ). Предлагаемая архитектура обладает следующими особенностями.

1. САО БМГД СБ является распределенной системой. Она охватывает: часть защищаемой КВИ; часть оборудования СКЦ; центральную часть (ядро) системы, которое находится на территории СПб ФИЦ РАН.

2. Компоненты 1–5 и 7 системы предназначены для решения конкретных задач по анализу событий безопасности и принятию решений по противодействию атакам. Компонент 6 предназначен для визуального анализа событий безопасности, а также представления промежуточных и конечных результатов работы предыдущих компонентов.

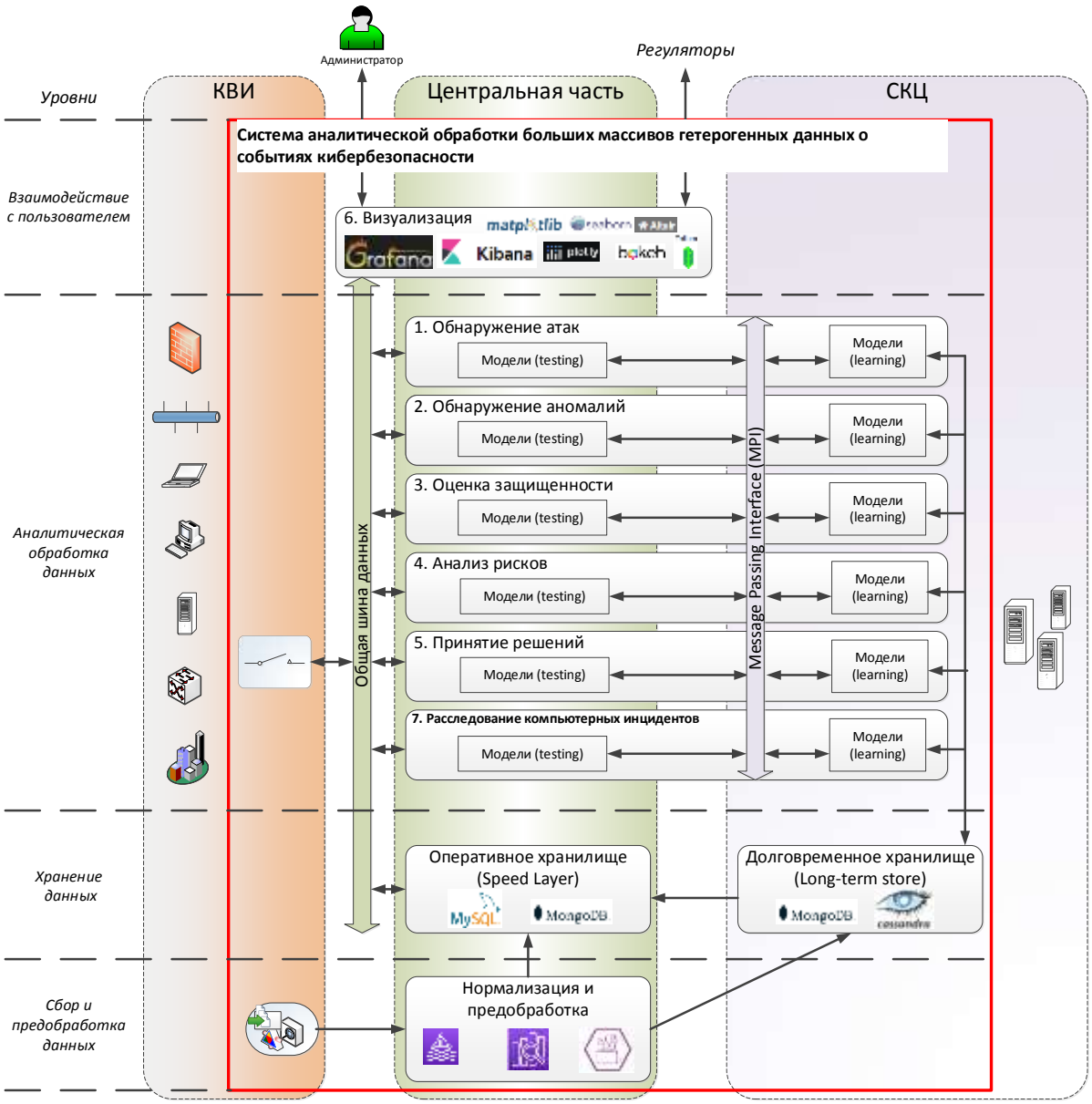

Рисунок. Архитектура системы аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности

3. Каждый из функциональных компонентов содержит 3 части оборудования: удаленную на стороне КВИ; удаленную на стороне СКЦ; центральную, находящуюся в СПБ ФИЦ РАН.

4. Сбор, предобработка и хранение данных о событиях безопасности происходят на уровнях «Сбора и предобработки данных» и «Хранения данных». На этих уровнях выделить два типа хранилищ данных: оперативное (Speed Layer), которое находится в центральной части, и долговременное (Long-term store), расположенное в СКЦ.

5. Обмен между компонентами осуществляется через общую шину данных. Кроме того, общая шина отвечает за доведение данных до администратора и регуляторов.

6. Каждый функциональный компонент состоит из двух частей: центральной и удаленной. В центральной части обученные модели применяются для тестирования (т. е. решаются задачи по выявлению атак, аномалий, оценки защищенности и т. д.). В удаленной части происходит обучение моделей. Обмен между центральными и удаленными частями компонентов происходит по технологии MPI.

Заключение. Таким образом, рассмотрены основные положения разработанной концепции аналитической обработки больших массивов гетерогенных данных о событиях кибербезопасности в интересах оценки состояния, поддержки принятия решений и расследования компью-

терных инцидентов в КВИ. Эти положения касаются подходов к применению суперкомпьютерных вычислений для обработки больших данных по кибербезопасности и архитектуры САО БМГД СБ. Направления дальнейших исследований связываются с разработкой моделей, методов, алгоритмов и прототипов программных средств аналитической обработки больших данных по кибербезопасности, предназначенных для реализации в соответствующих функциональных компонентах САО БМГД СБ.

Работа проводилась при поддержке гранта РНФ № 21-71-20078

ЛИТЕРАТУРА

1. **Yang X., Wu C., Lu K., Fang L., Zhang Y., Li S., Guo G., Du Y**. An Interface for Biomedical Big Data Processing on the Tianhe-2 Supercomputer. *Molecules.* 2017. Vol. 22, № 12. Article 2116.
2. **Demchenko Y., Ngo C., de Laat C., Membrey P., Gordijenko D**. Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. Lecture Notes in Computer Science. 2014. Vol. 8425. P. 76-94.
3. **Jain P., Gyanchandani M., Khare, N**. Big data privacy: a technological perspective and review. *J Big Data*. 2016. Vol. 3, № 25.
4. **Агеев С.А., Шерстюк Ю.М., Саенко И.Б., Полубелова О.В**. Концептуальные основы автоматизации управления защищенными мультисервисными сетями. *Проблемы информационной безопасности. Компьютерные системы*. 2011. № 3. С. 30-39.
5. **Kotenko I., Doynikova E**. Security Assessment of Computer Networks based on Attack Graphs and Security Events. *Lecture Notes in Computer Science*. 2014. Vol.8407. P.462-471.
6. **Браницкий А.А., Котенко И.В**. Обнаружение сетевых атак на основе комплексирования нейронных, иммунных и нейро-нечетких классификаторов. *Информационно-управляющие системы*. 2015. № 4 (77). С.69-77. doi:10.15217/issn1684-8853.2015.4.69.
7. **Mei S., Guan H., Wang Q**. An Overview on the Convergence of High Performance Computing and Big Data Processing. *Publication in ICPADS`18: Proceedings of the 2018 IEEE 24th International Conference on Parallel and Distributed Systems* (December 2018). **P.** 1046–1051.
8. **Десницкий В.А., Чечулин А.А., Котенко И.В., Левшун Д.С., Коломеец М.В**. Комбинированная методика проектирования защищенных встроенных устройств на примере системы охраны периметра. *Труды СПИИРАН*. 2016. № 5 (48). С. 5-31.
9. **Kotenko I., Konovalov A., Shorov A**. Agent-based simulation of cooperative defence against botnets. *Concurrency and Computation: Practice and Experience*, Vol. 24, Issue 6, 25 April 2012. P. 573-588.

I.V.Kotenko, I.B.Saenko (St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg)

The Concept of Analytical Processing of Large Arrays of Heterogeneous Data on Cyber Security Events in Critical Infrastructures

The concept of analytical processing of large arrays of heterogeneous data on cybersecurity events, produced in the interests of state assessment, decision support and investigation of computer incidents in critical infrastructures, is considered. Approaches to the use of supercomputing for processing big data on cybersecurity are analyzed. The architecture of a system for analytical processing of large arrays of heterogeneous data on cybersecurity events is proposed and discussed.

Авторы готовы представить текст на английском языке для сборника материалов мультиконференции, который будет подан для индексирования в Scopus