TPC[®] Newsletter

Issue 5, December 2024

TPC Technology Conference 2024 – TPCTC Guangzhou, China August 29th, 2024



TPCTC is TPC's annual technology conference. Its mission is to bring together industry experts and researchers to explore new methodologies for measuring the performance of datacentric applications. Over the last 15 years TPCTC has been recognized as the international event for anyone interested in performance related topics in database technology, including Transaction Processing, Data Warehousing, Big Data Analytics, Internet of Things, Virtualization, and Artificial Intelligence.

Performance evaluation has and still is one of the main differentiators of computer systems. Constant hardware and software improvements require a fresh look at performance methodologies that allow performance evaluation in a technically sound, fair and meaningful ways. The TPC has always been at the forefront of these benchmark developments. While the TPC has focused historically on database centric benchmarks, recent developments include benchmarks for Artificial Intelligence, Internet-Of-Things, Hyper-Converged Infrastructure, Big Data and Virtualization. Many of these benchmarks were sparked by ideas that originated in papers presented at past TPCTC events. These papers were mostly academic papers, that inspired ideas for new benchmarks, identified deficiencies in existing benchmarks and motivated improvements.

At the same time the academic community and industry have benefited from TPCTC to define their own performance methodologies. The last ten years have seen the rise of many new DBMSs, some with unique approaches to traditional problems, like columnar databases and in memory databases, others specializing on specific applications, such as graph databases, NoSQL databases, Timeseries databases etc. With them came a flurry of new performance methodologies resulting in customized benchmarks, many of which were based on methodologies originally developed in the TPC.

TPCTC has always served as a venue where both practitioners with real world performance expertise met with highly innovative academics to discuss performance methodologies for emerging technologies. The results, both in academia and TPC benchmarks, are a direct result of this knowledge exchange.

TPCTC 2024, chaired by Raghu Nambiar and Meikel Poess, took place in Guangzhou, China on August 29, 2024 in conjunction with the 50th conference on Very Large Databases (VLDB), a top tier academic conference for databases. TPCTC 2024 marked TPCTC's 16th anniversary.

This year TPCTC offered authors the option to submit the tools and scripts used to conduct performance measurements together with their paper. The TPC refers to the tools and scripts as a workload artifact.

Accepted workload artifacts will be hosted in TPC's GitHub repository and promoted on the TPC's website together with author names. The TPC will also provide a mechanism for results from workload artifacts to be listed on the TPC's webpage. The following accepted papers will be published in Springer's Lecture Notes in Computer Science (LNCS) and listed on ACM DL and DBLP.

PDSP-Bench: A Benchmarking System for Parallel and Distributed Stream Processing

Pratyush Agnihotri, Boris Koldehofe, Roman Heinrich, Carsten Binnig, and Manisha Luthra

Abstract: The paper introduces PDSP-Bench, a novel benchmarking system designed for a systematic understanding of performance of parallel stream processing in a distributed environment. Such an understanding is essential for determining how Stream Processing Systems (SPS) use operator parallelism and the available resources to process massive workloads of modern applications. Existing benchmarking systems focus on analyzing SPS using queries with sequential operator pipelines within a homogeneous centralized environment. Quite differently, PDSP-Bench emphasizes the aspects of parallel stream processing in a distributed heterogeneous environment and simultaneously allows the integration of machine learning models for SPS workloads. In our results, we benchmark a well-known SPS, Apache Flink, using parallel query structures derived from real-world applications and synthetic queries to show the capabilities of PDSP-Bench towards parallel stream processing. Moreover, we compare different learned cost models using generated SPS workloads on PDSP-Bench by showcasing their evaluations on model and training efficiency. We present key observations from our experiments using PDSP-Bench that highlight interesting trends given different query workloads, such as non-linearity and paradoxical effects of parallelism on the performance.

A Survey of Stream Processing System Benchmarks

Wang Yue, Martin Boissier, and Tilmann Rabl

Abstract: Stream processing systems are a fundamental component of modern data processing, enabling timely and efficient handling of streaming data. To assess and compare the capabilities of stream processing systems, various benchmarks have been proposed over the past years. Examples span a wide range of use cases, ranging from benchmarks for enterprise computing to social network analyses and IoT networks. These benchmarks are designed with different focuses and exhibit different characteristics during the execution. In this paper, we review existing stream processing benchmarks and analyze them with five dimensions: type, workloads, data ingestion, system under test (SUT), and metrics. We compare their similarities and differences, providing a comprehensive overview of existing benchmarks. Finally, we discuss the aspects that have been widely overlooked and highlight those that should be addressed when benchmarking future generations of streaming systems.

CrypQ: A Database Benchmark Based on Dynamic, Ever-Evolving Ethereum Data

Vincent Capol, Yuxi Liu, Haibo Xiu, and Jun Yang

Abstract: Modern database systems are expected to handle dynamic data whose characteristics may evolve over time. Many popular database benchmarks are limited in their ability to evaluate this dynamic aspect of the database systems. Those that use synthetic data generators often fail to capture the complexity and unpredictable nature of real data, while most real-world datasets are static and difficult to create high-volume, realistic updates for. This paper introduces CrypQ, a database benchmark leveraging dynamic, public Ethereum blockchain data. CrypQ offers a high-volume, ever-evolving dataset reflecting the unpredictable nature of a real and active cryptocurrency market. We detail CrypQ's schema, procedures for creating data snapshots and update sequences, and a suite of relevant SQL queries. As an example, we demonstrate CrypQ's utility in evaluating cost-based query optimizers on complex, evolving data distributions with real-world skewness and dependencies.

StarBench: A Fresh Approach On Star Schema Benchmarking

Hanumath Rao Maduri, Ahmad Ghazal, Alain Crolotte and Yuchao Li Hamish Nicholson, Andreea Nica, Aunn Raza, Viktor Sanca, and Anastasia Ailamaki

Abstract: In OLAP (Online Analytical Processing) environments, star schema outperforms 3NF (Third Normal Form). It offers simpler querying and superior performance, along with better tool support, making it analyst-preferred. Currently, the SSB benchmark is the main star schema benchmark but it has limited workload compared to other OLAP benchmarks like TPC-H. This paper introduces StarBench, a novel star schema benchmark that combines the advantages of star schema data model with the comprehensive workload of TPC-H. StarBench utilizes a 4-table denormalized model derived from TPC-H, effectively addressing the data gaps present in SSB. Data loading in StarBench is performed via insert-select operations from TPC-H tables. The workload remains consistent with TPC-H, incorporating SQL modifications to align with the new data model. We evaluated StarBench on PostgreSQL to demonstrate the benchmark's feasibility.

Web3Bench: A Web3 Based HTAP Benchmark

Ahmad Ghazal, Zhongxin Ge, Hanumath Rao Maduri, Anita Shao, Guoxin Kang, Jingpei Hu, Huaiyu Xu, Ruoxi Sun, Li Shen, and Ed Huang

Abstract: Hybrid Transaction/Analytical Processing (HTAP) databases are gaining popularity in the database community. The growing prominence of HTAP databases necessitates the development of suitable benchmarks tailored to evaluate these systems effectively. We believe existing HTAP benchmarks are not representative, as most of them merged the existing benchmarks for transactional and analytical workloads. Our research also found that the latest proposals did not take into account the online aspect of an HTAP database. Most of the existing proposals re-used previous benchmark data models that are not truly representative of HTAP workloads. This paper introduces Web3Bench, an HTAP benchmark that addresses the limitations mentioned above in the existing HTAP benchmarks. The benchmark we propose in this paper is based on a realistic workload in the Web3 domain, considered a predominant use case for an HTAP database. Our data model is a simplified version of the decentralized blockchain Ethereum dataset. In this benchmark, we leverage a sample data set from Ethereum to build a scale factor-based data generator. The workload in Web3Bench focuses on simple queries representing online processing with a high number of queries per second. In this paper, we also present the outcomes of our proof of concept conducted with this benchmark on the TiDB cluster, a distributed SQL database with HTAP features.

Evaluation Considerations of Synthetic Natural Language Datasets for Question Answering Applications

Chris Van Buren, Ajay Dholakia, Xiaotong Jiang, David Ellison, Sachin Gopal Wani, and Jieyu Lin

Abstract: The rapid advancements in transformer-based language models have opened new possibilities for applications such as question answering and synthetic dataset generation. However, generating highquality synthetic datasets can be computationally expensive and require significant resources. Fine-tuning a language model on a synthetic dataset can lead to improved performance on context-dependent questions, but it is crucial to evaluate the quality of these datasets beforehand. This study explores the benchmarking of synthetic natural language datasets generated with open-source language models for question answering applications. We fine-tune the same base model on each dataset independently and assess its performance on common benchmarks for question answering with context. Additionally, we examine metrics of the synthetic datasets themselves to identify potential indicators of downstream benchmark performance. Our results show that measuring the semantic similarity between questions is an indicator of domain diversity in a synthetic question-and-answer dataset. We also demonstrate that short answer lengths may indicate low quality chain-of-thought answers. Overall, Llama 3 8B Instruct performs most consistently well, of the models considered, in synthetic dataset generation. Our findings provide valuable insights for practitioners seeking to develop effective synthetic datasets for language based applications. In addition to the above papers this year's TPCTC featured a keynote by FeiFei Li, an invited talk by Pengcheng Zheng and a panel moderated by Pengcheng Zheng.

Keynote: Evolving the TPC benchmarks for Cloud Native Databases, presented by FeiFei Li



Feifei Li received the BS degree in computer engineering from Nanyang Technological University, Singapore, in 2002, and the PhD degree in computer science from Boston University, Boston, Massachusetts, in 2007. He is currently a vice president of Alibaba Group, ACM distinguished scientist, president of the Database Products Business Unit, Alibaba Cloud Intelligence, and director of the Database and Storage Lab, DAMO Academy

Invited Talk: Performance Evaluation of TimechoDB using TPCx-I was authored by Xinhao Gu, Xinyu Tan, Jialin Qiao, Junzhi Peng, Steve Yurong Su, Pengcheng Zheng, Xiangdong Huang, Shaoxu Song, and Jianmin Wang.

Abstact: The Internet of Things (IoT) is revolutionizing industries by generating an unprecedented volume of time series data, making robust and high-performance time series databases crucial for effective data management and analysis. Time series data is pivotal in sectors such as energy, finance, and manufacturing, each with unique challenges and requirements, which complicates the use of a universal benchmark for evaluating databases in different domain applications. This paper discusses benchmarking methodologies in evaluating time series databases, with a focus on IoT environments. Key characteristics like high cardinality for handling IoT data are highlighted, along with the integration of Artificial Intelligence (AI) technologies for analytics and the applicability of the TPCx-IoT benchmark. Future improvements to benchmarks are also considered to address the evolving demands of IoT applications.

Panel: Benchmarking Timeseries Databases: Current State and Future Perspectives, moderated by Pengcheng Zheng, VP at Alibaba Group

Panelists: Prof. Jianmin Wang - Tsinghua University, Raghu Nambiar - AMD, General Chair of TPCTC, Prof. Lei Chen - HKUST, VLDB 2024 General Chair, Prof. Hongzhi Wang - Harbin Institute of Technology, Prof. Mingsheng Long - Tsinghua University, Qiang Li - CISDI Info and Pengcheng Zheng - Timecho. See short bios below.

Prof. Jianmin Wang - Tsinghua University



Prof. Jianmin Wang is the Dean of the School of Software, Tsinghua University, and the Executive Director of National Engineering Research Center for Big Data Software. His research interests include unstructured data management, workflow and BPM technology, and database system. Over the years, he has made significant contributions to the field, authoring numerous influential papers and leading critical research projects. His work has garnered several accolades, including the National Science and Technology Progress Award. He has published over 100 DBLP indexed papers in Journals, such as TKDE, TSC, DMKD, CII, DKE, FGCS, and IJIIS, and in conferences, such as VLDB, SIGMOD, SIGIR, ICDE, AAAI, IJCAI, ICWS, and SAC. Beyond his academic achievements, Professor Wang plays a vital role in guiding the next generation of scholars and engineers, shaping the future of technology and software development in China and globally.

Raghu Nambiar - AMD, General Chair of TPCTC



Raghu Nambiar is a Corporate Vice President at AMD where he leads a global engineering team responsible for the strategy, roadmap, and execution for AMD's datacenter business. Prior to joining AMD, as the CTO of the Cisco UCS business he played an instrumental role in accelerating the growth of Cisco UCS to a top datacenter-compute platform. At Hewlett-Packard, where Raghu spent his early years as a technology leader, he was responsible for several industry-first and disruptive technology solutions and a decade of performance benchmark leadership. He has published more than 75 peer-reviewed papers and book chapters, 15 books in Lecture Series in Computer Science (LNCS) and holds ten patents with several pending. He holds dual Master's degrees from University of Massachusetts

and Goa University, and an advanced management program from Stanford University.

Prof. Lei Chen - HKUST, VLDB 2024 General Chair



Prof. Lei Chen is a Chair Professor in the Data Science and Analytic Thrust at HKUST (GZ), Fellow of the IEEE, a Distinguished Member of the ACM, and General Chair of the VLDB 2024. Currently, Prof. Chen serves as the Dean of Information Hub, the Director of Big Data Institute at HKUST, MOE/MSRA Information Technology Key Laboratory. His research interests include Data-driven AI, knowledge graphs, blockchains, data privacy, crowdsourcing, spatial and temporal databases and query optimization on large graphs and probabilistic databases. Prof. Chen received the SIGMOD Test-of-Time Award in 2015, Best research paper award in VLDB 2022. The system developed by his team won the excellent demonstration award in VLDB 2014. Prof. Chen had served as VLDB 2019 PC Co-chair. Currently, Prof. Chen serves as Editor-in-chief of IEEE Transaction on Data and

Knowledge Engineering and an executive member of the VLDB endowment.

Prof. Hongzhi Wang - Harbin Institute of Technology



Prof. Hongzhi Wang is a distinguished professor and Ph.D. advisor at Harbin Institute of Technology (HIT). He serves as the Head of the Department of Computer Science and Engineering, Director of the Center for Mass Data Computing, and is responsible for the Data Science and Big Data Technology program. He also leads the Heilongjiang Province Key Laboratory of Big Data Science and Engineering and the HIT Youth Scientist Studio. Prof. Wang is a Distinguished Member of the China Computer Federation and a Senior Member of IEEE. His research focuses on databases, big data management and analysis, and big data

governance. He has published over 350 papers, with more than 100 indexed by SCI and cited over 4,000 times, and has led more than 10 projects, including National Natural Science Foundation key projects and international collaborations.

Prof. Mingsheng Long - Tsinghua University



Qiang Li - CISDI Info

Prof. Mingsheng Long is the Director of Institute of Software System Engineering in the School of Software at Tsinghua University. His research spans machine learning theory, algorithms and models, with persistent dedication to creating strong learning machines from big data that adapt to complex real world. He has published a long article in Nature and received a feature report, was featured on the cover of Nature Machine Intelligence, and won the Test of Time Award of IJCAI. Prof. Long received the BE and PhD degrees from Tsinghua University, in 2008 and 2014 respectively. He was a visiting researcher with UC Berkeley from 2014 to 2015. He serves as an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and the Artificial Intelligence Journal, and as (Senior) Area Chairs of major machine learning conferences, including ICML, NeurIPS, and ICLR



Qiang Li is the Deputy General Manager and CTO of CISDI Info, a distinguished talent under the "Golden Phoenix" program in the Western (Chongqing) Science City, Deputy Director of the Chongqing Key Laboratory of Industrial Software and Cloud Innovation, an expert in intelligent manufacturing of the China Iron and Steel Association (CISA), and Vice Chairman of the Chongqing Copyright Association. He has led the development of "CISDigital IIoT Platform", which won the national championship of the China Industrial Internet Contest, and recognized by the Ministry of Industry and Information Technology (MIIT) as a national-

level "Cross-industry and Cross-regional" platform. The platform serves 18 industries and has established in-depth cooperation with more than 100 benchmark enterprises.

Pengcheng Zheng - Timecho



Pengcheng Zheng is the Managing Director of Timecho Europe, concentrating on time series database management system and related technologies to address challenges in data management in the era of IoT. Pengcheng previously worked at the Fraunhofer Institute for Industrial Engineering in Germany, where he participated in several EU research projects and contributed to multiple publications, with a focus on Industry 4.0 and automotive industry research. Additionally, he is dedicated to open-source software advocacy, particularly focusing on Apache IoTDB and PLC4X. At Timecho, his team maintains close collaboration with academic institutions like Tsinghua University and industry partners to establish a de facto standard for next-generation time series databases

TPCTC 2024 Organization

General Chairs and Contacts

Raghunath Nambiar, AMD, USA, <u>raghu.nambiar@amd.com</u> Meikel Poess, Oracle, USA, <u>meikel.poess@oracle.com</u>

Program Committee

Ajay Dholakia, Lenovo, USA Andrew Bond, Red Hat, USA Anil Rajput AMD, USA Hans-Arno Jacobsen, University of Toronto, Canada Harry Le, University of Houston, USA John Poelman, IBM, USA Klaus-Dieter Lange, Hewlett Packard Enterprise, USA Michael Brey, Oracle, USA Miro Hodak, AMD, USA Nicholas Wakou, Dell, USA Paul Cao, Hewlett Packard Enterprise, USA Rodrigo D. Escobar, Univ. Texas at San Antonio, USA Shahram Ghandeharizadeh, University of Southern California, USA Tariq Magdon-Ismail, VMware, USA Tilmann Rabl, Hasso Plattner Institute, Germany