PostgreSQL数据库的智能存储

**KaiGai Kohei The PG-Strom Development Team**

出自：<http://heterodb.com/blobs/P7130_KAIGAI_SSD2GPU_FIXTYPO.pdf>

翻译： PostgreSQL中文社区**朱君鹏**

# 文档声明

所有版权归原作者所有，本文只是负责翻译和分享工作，如果需要引用标明出处。

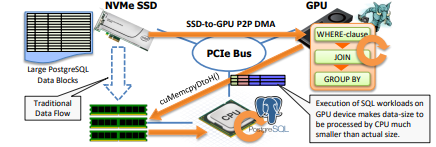
# 修改记录和版本信息

| 日期 | 作者 | 版本 | 修改记录 |
| --- | --- | --- | --- |
|  |  |  |  |
| 2019-2-23 | **朱君鹏** | 1.0 | 创建初始版本 |
| 2019-04-22 | Qinghui.guo | 2.0 | 添加版本信息 |
| 2019-08-10 | 崔鹏 | 3.0 | 文章校对 |
|  |  |  |  |

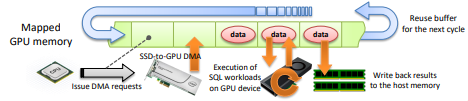
# 摘要

GPUDirect RDMA允许直接从PCIe设备到GPU RAM的对等数据加载。对于Linux内核和PostgreSQL的扩展模块，我们通过将NVMe-SSD上的数据库块加载到GPU RAM以及在GPU设备上执行SQL来协同利用此基础架构进行非常快速的表扫描。一旦数据块加载到GPU RAM上，内核函数就会根据提供的SQL（WHERE clause，JOIN和GROUP BY）减少数据大小。在结果中，CPU / RAM将获得比实际表大小小得多的数据大小，并且看起来存储在理解SQL的情况下智能地执行。根据基于SQL星型模式的基准测试，我们的功能可以在80秒内扫描351GB平台;这是大约4.5GB / s的查询处理吞吐量，比通常的文件系统基本I / O实现快2.5倍。此结果表明GPU对I / O密集型工作负载也很有价值，而不仅仅是计算密集型工作负载。

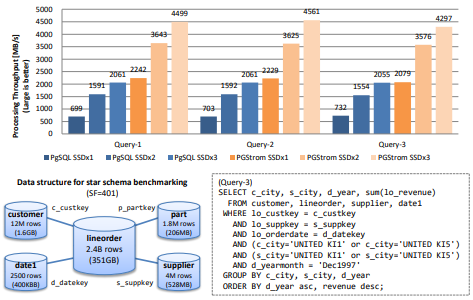
# 简要架构



OLAP（联机分析处理）是数据库领域的一个主要工作。它的特点是查询想要从大量数据中生成（相对）小的摘要;通常大于物理RAM大小。传统上，存储设备（如SSD）上的数据库的数据块被加载到主机系统的RAM一次，然后CPU或GPU处理它们。但是，由于数据块中的大多数行都被过滤，所以效率不高，因为在GROUP BY操作的缩减过程中，数据块中的大多数行都被过滤掉或仅被引用一次。在SQL工作负载的情况下，我们可以准确地知道如何在加载之后处理数据块。因此，可以设计GPU内核以并行地在多个块/行上运行SQL工作负载。我们实现了WHERE子句评估的GPU版本，Hash-Join和GROUP BY，作为PG-Strom [\* 1]的功能。此外，我们开发了一个Linux内核模块，用于将SSD从SSD块传输到GPU RAM。PG-Strom是PostgreSQL [\* 2]数据库的扩展。它控制Linux内核模块和GPU设备以卸载SQL工作负载。该架构允许一些SSD和GPU协同工作，并根据提供的SQL将预处理的数据流发送回主机系统。从应用程序的角度来看，存储似乎可以智能化的在设备上运行一些SQL工作负载。使用GPU RAM（=相对罕见）处理大数据的关键是循环缓冲和异步执行。我们将映射的GPU内存分成多个块，并逐个分配作业。在特定时刻，CPU在块上发出SSD到GPU DMA请求，在另一个块上的请求正在进行中，GPU内核正在运行以处理另一个块，并且通常的GPU到RAM写回DMA在 - 另一个块上的进展。它可以非常有效地提取GPU和NVMe-SSD的硬件功能。



# 基准测试



我们在原始NVMe-SSD [\* 3]上构建了一个具有上述数据结构的数据库，以及由2或3个SSD设备组成的md-raid0卷。此工作负载是众所周知的星型模式基准，它反映了典型的DWH工作负载，也类似于一些IoT类数据结构。我们运行一些报告查询，如上面的Query-3。最大的表大小为351GB，但RAM大小为128GB。因此，此工作负载是典型的I / O密集型工作负载而非计算密集型。蓝色条是每个RAID配置的vanilla PostgreSQL的结果，橙色条是PG-Strom的结果，具有SSD到GPU的P2P DMA支持。查询处理时间的倒数（例如，如果查询在80秒内完成，则其数据处理为351GB / 80.0s = 4492MB / s）。它显示了更多的SSD设备使数据处理吞吐量更高。另一方面，PG-Strom的结果显示，SSD-to-GPU P2P DMA比基于文件系统+ CPU的数据处理能够提供更高的NVMe-SSD功能。特别是，“PG-Strom SSDx1”是原始NVMe-SSD设备的结果。其数据处理吞吐量约为2.1-2.2GB / s;这几乎相当于SSD设备的目录规范（SeqRead中为2.2GB / s）。这些结果向我们介绍了最近的半导体创新（GPU，SSD，......）使我们能够以比以前更便宜的成本赶上高端DWH级性能。

# 结论和展望

我们的挑战表明，从应用的角度来看，一对GPU和NVMe-SSD可以像智能存储一样运行。 GPU可以根据应用程序端的知识处理数据流，但在数据到达CPU / RAM之前有GPUDirect RDMA支持。它将数据处理吞吐量提高到了硬件的理论极限值附近，但成本比现有的高端数据库解决方案相对较小。目前，我们的实现仍然在md-raid0配置上具有不可忽视的开销。它需要修改。然后，我们将尝试每个计算节点10GB / s的数据处理吞吐量。

# 参考文献

[\*1] PG-Strom – Extension of PostgreSQL for GPU acceleration, http://strom.kaigai.gr.jp/

[\*2] PostgreSQL – Well used open source RDBMS, <https://www.postgresql.org/>

[\*3] Intel SSD 750 (400GB) – http://www.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-750-series.html