Cloud Workload Characterization

Wira D. Mulia, Naresh Sehgal, Sohum Sohoni, John M. Acken, C. Lucas Stanberry and David J. Fritz
Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK,
Platform Enabling Group, Intel Corporation, Santa Clara, CA,
Engineering and Computing Systems, Arizona State University, Mesa, AZ, USA

Contribution

--describes the Cloud workload categories as differentiated by the customer and vendor viewpoints, connects these workload categories to their computer resource requirements

-- describes the low-level hardware measurements that can be used to distinguish job transitions between categories and within phases of categories

-- low-level hardware measurements can also be used to detect resource contention between the workloads and, in conjunction with the categories, minimize contention to improve resource allocation

-- SLAs(service-level agreements), supplied capital purchase decisions, and future computer architecture design decisions can be made based upon the categories described in this paper

Steps to tackle the Cloud workload categorization

-- the first step is to define a comprehensive list of the computer workload categories
-- The second step is to identify all of the resources required for Cloud Computing
-- The third step is to associate each Cloud Computing workload category with the key, largest, and performance-limiting computer resources.
-- The fourth step is to correlate low-level hardware metering with the Cloud Computing categories

Workload categories

Top-level Categorization

--Static architecture
  -- the implementation of solution architecture such as big data storage or a parallel computational setup

--Dynamic behavior
  -- the run-time nature of resource usage, or stress that a workload places on the computational resources in a DC

Cloud Workload Categories

lists and briefly describes the categories of workloads as differentiated by the customer and vendor viewpoints

-- The perspective in this paper is to define the categories based upon fundamental concepts

-- categories are definitely linked to and related to underlying computer resources; however, they are not defined by those resources.

--This means the categories remain even when the implementation hurdles are removed. Another categorization hurdle is the dependence upon the user's perspective

--categories incorporate the user's perspective

--these categories are an operational viewpoint of Cloud Computing workloads

--The categories of computing workloads considered in this section include both Cloud computer workloads and workloads that are not commonly considered Cloud Computing issues

Cloud Workload Categories

--big streaming data workload category
  --characterized by an interactive initiation followed by long periods of huge amounts of data (usually video) sent to the end customer
  --This workload is usually measured by data throughput rather than latency
  --Therefore, the key underlying hardware resources are the bandwidth to the user, the server delivery, and the network speed from the storage to the server

--big database creation and calculation category
  --characterized by a large amount of data requiring simple but massive computation efforts

--big database search and access workload category
  --characterized by repeated interactive requests or queries submitted to a very large database
  --The customer satisfaction for this category is dependent upon the response time or latency
  --The key resources that limit the latency are the database server software and organization, the disk storage, the network bandwidth from the storage to the database server, and the server load

--big data storage workload category
  --characterized by the integrity of large amounts of data which is periodically updated, usually by small increments, and that occasionally requires a massive download or update
  --The speed (latency, throughput, storage, incremental processing, and complete restoration) is not the highest priority

--in-memory database workload category
  --characterized by the integrity of large amounts of data which is rapidly accessed
  --The limitation of this category is primarily size of the data
  -- The speed (latency, throughput, storage, incremental processing, and complete restoration) is the highest priority

--many tiny tasks (ants) workload category
  --characterized by several very small tasks running independently
  --This category depends upon the ability to assign multiple processors

--tightly coupled intensive calculation HPC workload category
-- characterized by problems requiring teraflops of computing power

--separable calculation 0 intensive HPC workload
--characterized by large time-consuming quantities of calculations
--characterized by the ability to split up the calculations to be performed in parallel or small separable pieces

--highly interactive single-person tasks workload category
--characterized by a fast task with a single user.
--The key aspect here is response time latency

-- highly interactive multi-person jobs workload category
--characterized by high-speed, large, single-user tasks that have significant user interaction
--All of the computation for one designer’s part is local, but the database for the overall chip is on the network or Cloud

--private local tasks workload category
  --characterized by traditional single-user tasks

--slow communication workload category
  --characterized by small amounts of information without a time limit for delivery

--real-time local tasks workload category
  --characterized by hardware measurements fed to a computer system
  --The resource key here is dedicated CPUs with interrupt-driven communication to the network

-- location aware workload category
  --characterized by utilizing auxiliary location input data (such as GPS or telephone area code) to small amounts of information without a time limit for delivery

--real-time geographically dispersed tasks workload category
  --characterized by several dispersed hardware measurement systems feeding data to a network
  --The resource key here is dispersed dedicated CPUs with measurement and interrupt-driven communication to the network servers

--access control workload category
  --characterized by user-initiated requests where the response is to another server authorizing more activity
  --The access control workload category must balance the tradeoff between Cloud provider security and Cloud user privacy

--voice or video over IP workload category
  --characterized by user-initiated requests where the response is through a server to each other
  --The resources required are network bandwidth and user local compute power for data compression and decompression

Computing Resources

--connecting low-level resources to Cloud workload categories
--All metrics listed are potentially used by Cloud providers
  --Persistent storage
  --Compute power/computational capability, metrics related to CPU
  --Network bandwidth
  --Broadcast transmission receivers
  --Data busses within a server such as CPU to memory, cache to main memory, memory to disk, and backplanes on a card

Table 1: Characteristic computing resources for workload categories for the Cloud

Workload category	User view or example providers	Limiting resources	Level of Cloud relevance: "How Cloud heavy is this category?"
Big streaming data	Netflix	Network bandwidth	Heavy
Big database creation and calculation	Google, US Census	Persistent storage, computational capability, caching	Heavy
Big database search and access	US Census, Google, online shopping, online reservations	Persistent storage, network, caching	Heavy
Big data storage	Rackspace, Softlayer Livedrive, Zip Cloud Sugarsync, MyPC	Persistent storage, caching, bus speed	Heavy
In-memory database	Redis, SAP HANA, Oracle In-Memory DB	Main memory size, caching	Heavy
Many tiny tasks (ants)	Simple games, word or phrase translators, dictionary	Network, many processors	Heavy
Tightly coupled calculation-intensive HPC	Large numerical modeling	Processor speed, processor to processor communication	Medium
Separable calculation-intensive HPC	CCI on Amazon Web Services, Cyclone™ (SGI) Large simulations	Processor assignment and computational capability	Heavy
Highly interactive single person	Terminal access, server administration, web browsing, single-player online gaming	Network (latency)	Some
Highly interactive multi-person jobs	Collaborative online environment, e.g., Google Docs, Facebook, online forums, online multiplayer gaming	Network (latency), processor Assignments (for VMs)	Medium
Single computer intensive jobs	EDA tools (logic simulation, circuit simulation, board layout)	Computational capability	None
Private local tasks	Offline tasks	Persistent storage	None
Slow communication	E-mail, blog	Network, cache (swapping jobs)	Some
Real-time local tasks	Any home security system	Network	None
Location aware computing	Travel guidance	Local input hardware ports	Varies
Real-time geographically dispersed	Remote machinery or vehicle control	Network	Light now, but may change in the future
Access control	PayPal	Network	Some, light
Voice or video over IP	Skype, SIP, Google Hangout	Network	Varies

--Table 1 relates the workload categories to computer system resources

--The table includes computing jobs that are not Cloud related for two reasons.

--Primarily the non-Cloud categories are included for completeness.

--Secondly, in the future, all computing may be in the Cloud

Temporal Variability of Workloads

-- two different cases in which the workload category would change
  -- The first case is when the next step or phase of a job is a different category than the current category
  -- The second case is when the job is incorrectly categorized

Low-level or Hardware Metrics of Computer Utilization

--instruction per cycle (IPC)
  -- a good indicator of whether the CPU is being fully utilized or not

-- last level cache (LLC) misses
  -- gives us an idea of how much memory traffic is being generated

-- L1 Data Cache Misses
  -- L1 data misses allows us to gauge whether the core working set of an applications is captured in the first-level cache

-- Cycles for Which Instruction Queue is Full
  -- this could indicate that the pipeline is functioning at a high level of utilization if the IPC is high (which is good)

-- L1 Instruction Cache Misses
  -- a large number of misses indicates an issue with the front-end and the instruction fetch mechanism

-- Cycles During Which Reservation Stations Are Full
  -- By observing specific reservation stations, and with some knowledge about the typical usage characteristics of applications, we can identify the applications running on a particular server

-- Number of Lines Fetched from Memory
  -- estimate of the pressure on the main memory

Conclusion

-- identifies the relationship of critical computer resources to various workload categories.
--Therefore, low-level hardware measurements can be used to distinguish job transitions between categories and within phases of categories