SUMMIT2017

PASS

Monitoring Linux Performance for the SQL Server Admin

Anthony E. Nocentino, Enterprise Architect, Centino Systems

Anthony E. Nocentino

Enterprise Architect, Centino Systems
aen@centinosystems.com

in /nocentino

@nocentino

Consultant and Trainer

Founder and President of Centino Systems

Specialize in system architecture and performance

Computer Science, M.S. and B.S.

Microsoft MVP - Data Platform

Friend of Redgate

Linux Foundation Certified Engineer

Microsoft Certified Professional

Other places online…

Blog - www.centinosystems.com/blog

Pluralsight Author

Centino systems

Agenda

- Linux System Architecture
- SQL on Linux Architecture
- System Components
    - CPU/Processes
    - Memory/Pages
    - Disk/File Systems
- Monitoring Tools

Things we’re going to cover

- Linux OS concepts, how it works!
- Tools to view performance data
- What’s good and what’s bad

Things we’re NOT going to cover

- SQL Server internals
- Performance troubleshooting

Linux Architecture

User Space	Users	Interact with the Shell	Cause Problems :)
	Shell	Executes Your Commands…Your Interface to the Kernel	Commands, Editors…any User Program
Kernel Space	Kernel	Resource Management and Access	Process, Pages and File Systems
	Hardware	Physical Resources	CPU, Memory and Disk

SQLOS

Scheduling

Placing tasks into workers and getting access to the CPU

Synchronization

Controlling access to system resources

I/O

Scheduling of I/O both network and disk

Memory Management

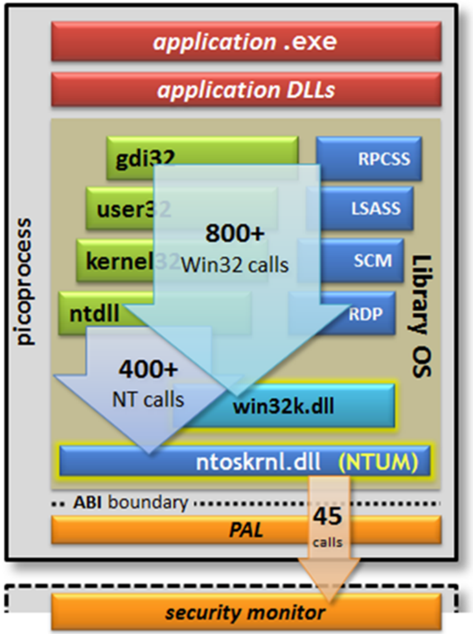
Allocation of memory to various system objects

Primary function is resource management specific to RDBMS

“A new platform layer in SQL Server 2005 to exploit new hardware capabilities and their trends” S. Oks

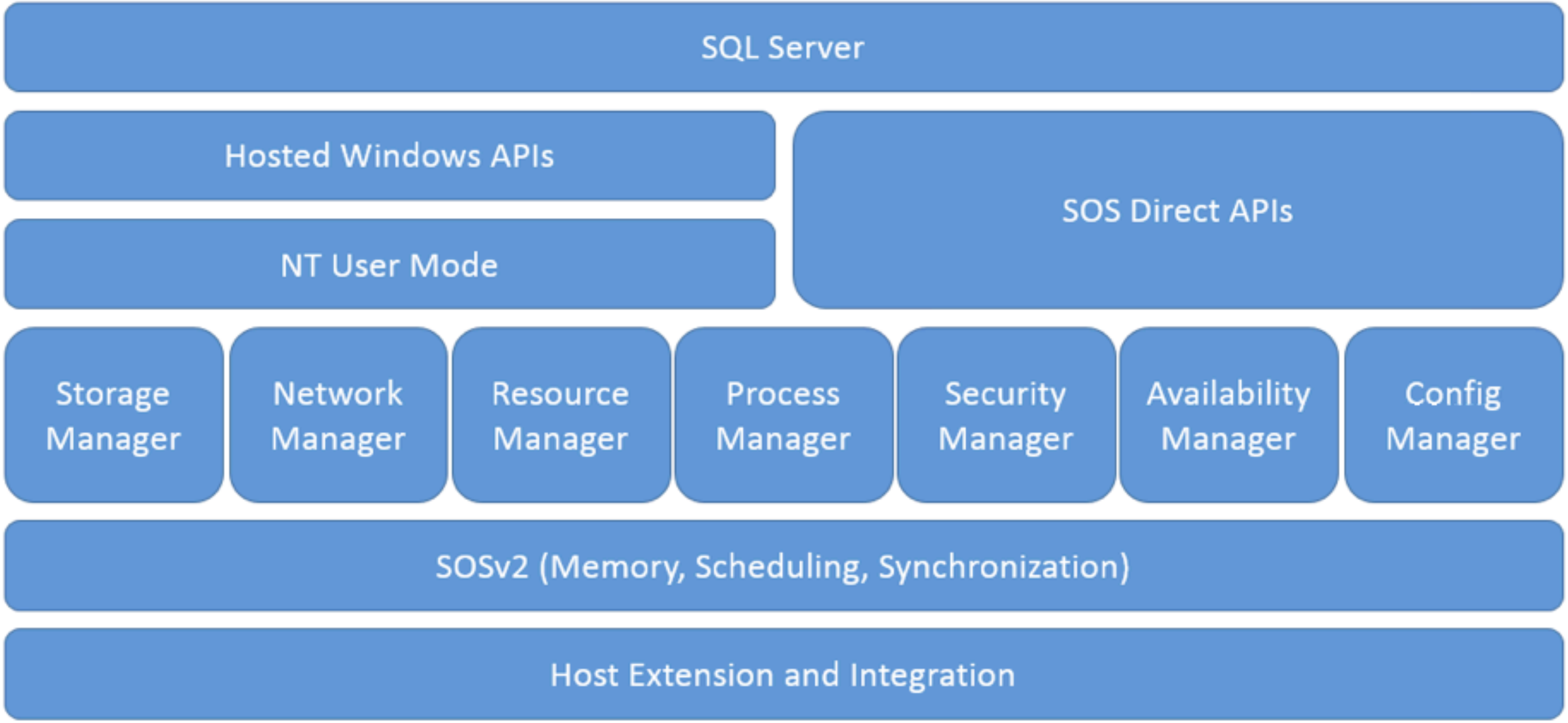
“Operating System support for Database Management” M. Stonebraker

SQL on Linux Architecture - Drawbridge



From: https://blogs.technet.microsoft.com/dataplatforminsider/2016/12/16/sql-server-on-linux-how-introduction/

SQL on Linux Architecture - SQLPAL



SQL Server

Hosted Windows APIs

NT User Mode

SOS Direct APIs

Storage Manager

Network Manager

Resource Manager

Process Manager

Security Manager

Availability Manager

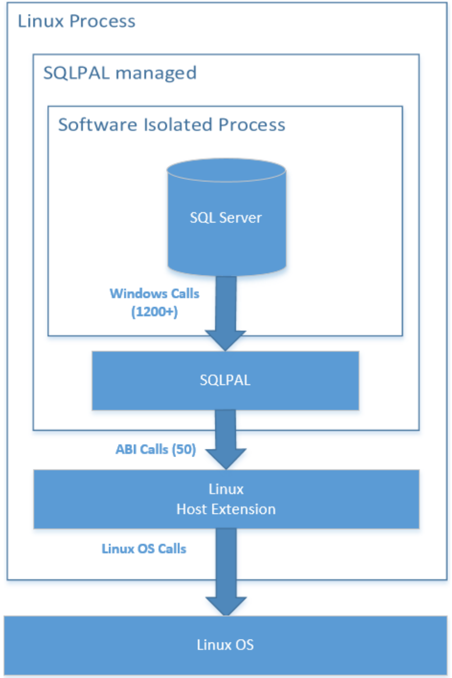
Config Manager

SOSv2 (Memory, Scheduling, Synchronization)

Host Extension and Integration

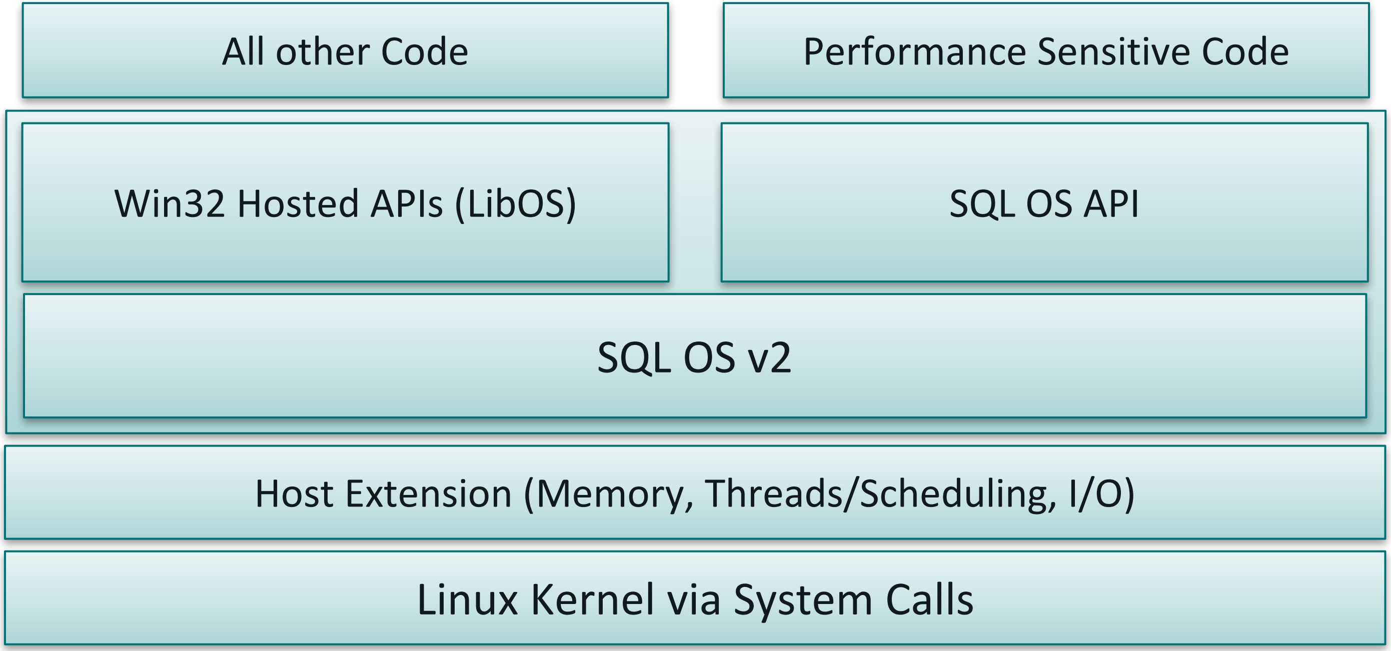
From: https://blogs.technet.microsoft.com/dataplatforminsider/2016/12/16/sql-server-on-linux-how-introduction/

SQL on Linux Architecture - Process Layout



From: https://blogs.technet.microsoft.com/dataplatforminsider/2016/12/16/sql-server-on-linux-how-introduction/

SQL on Linux Architecture - SQLPAL


All other Code
Performance Sensitive Code
Win32 Hosted APIs (LibOS)
SQL OS API
SQL OS v2
Host Extension (Memory, Threads/Scheduling, I/O)
Linux Kernel via System Calls

From: https://blogs.technet.microsoft.com/dataplatforminsider/2016/12/16/sql-server-on-linux-how-introduction/

SQL on Linux Architecture - Host Extensions

- Call table maps Win32 API semantics to Linux System calls
- ~45 ABI Calls
  - Memory Management
  - Threads and Scheduling
  - Synchronization Primitives
  - I/O Network and Disk
- We care a lot about host extensions…it’s more code

Shhhhhh - SQLPAL is Virtualization ;)

- **Process virtualization (not machine)**
  - Presenting another environment inside the process’ context that’s different than that of the hardware’s operating environment

- But the environment is purpose built for SQL Server

- We need to understand that this is a hybrid Win32/Linux process and have a firm grasp of
  - Resource allocation and management in SQLPAL
  - How that turns into Linux OS performance
  - Debugging

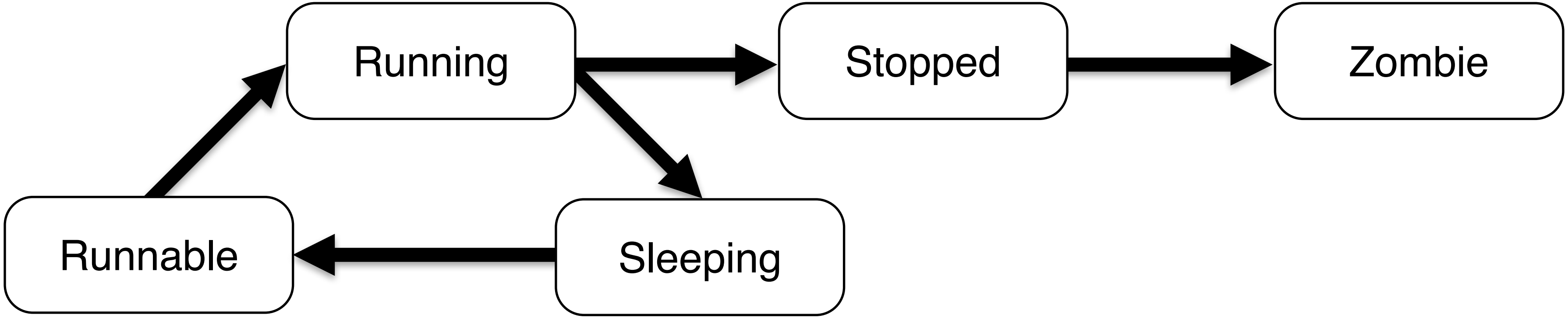
CPU and Processes

What is a Process

- Process
  - Executing program, program code, memory and resources
- Thread (LWP)
  - Shared access to resources
- Process and Thread Creation
  - `fork/exec` - parent process yields a child process with a PID
  - `clone` - same address space as thread creator, cheap and fast!
- Process Tree
  - The hierarchy of parent and it’s child processes

What is a Process (con’t)

- Process States



Controlling Processes

- Signals
- Methods of process control
  - `kill` and `killall`
- Niceness
- Set the execution priority
  - `nice` and `renice`
  - Default 20, lower is less “nice”

More on Processes…

- Context switching
- Kernel versus User Mode
- CPU Scheduling
    - How is a SQLOS Worker scheduled onto the CPU?
        - Creates a thread via pthread and that’s pushed into the scheduler
    - pthreads?

Process/Thread Scheduling

- Unit of scheduling is the thread
- Default scheduler is `SCHED_OTHER/SCHED_NORMAL`

- Time sharing scheduler
  - Preemptive
  - Dynamic priority list, based on niceness
  - Calculated quantum length based on priority
    - `kernel.sched_min_granularity_ns = 10000000` (10ms) - default
    - `kernel.sched_wakeup_granularity_ns = 15000000` (15ms) - default
  - NUMA Aware, but…
    - `kernel.numa_balancing = 0 -` default

CPU - What to look for?

- Percentage of what?
- Load average
- Run queue length and I/O waits
- Spikes aren’t bad
- Long waits
  - User
  - I/O - disk latency will effect access to the CPU
  - System

Tools to use for process monitoring

- `top/htop`
- `ps`
- `mpstat/pidstat`
- `dstat`
- `procfs`

Demos

- Processes and threads
- Run load average under CPU saturation
- Exploring `procfs`

Memory and Pages

Memory

- Memory Layout and Architecture
  - Physical and Virtual Memory
  - NUMA - free lists per node
  - Pages (Anonymous)
  - Demand Paging
    - Swap out
      - Time and Pressure
    - Swap in, Major Page Fault
    - Allocation, Minor Page Fault
  - File System Cache and swappiness - http://red.ht/2cHg9Vk
    - `vm.swappiness = 10` (default 30, 0 disables swapping)
    - `vm.dirty_ratio = 40 (default 30)`
    - `vm.max_map_count = 262144` (default 65530)

Pages

- Regular pages - 4KB
- Transparent huge pages - 2MB
    - Increases memory I/O by decreasing TLB cache misses
- SQLOSv2
    - Can request large pages inside SQL Server…with trace flag 834
        - SQL will allocate memory on start up
        - When SQLPAL exposes 8GB+ to SQL Server
- As of today, no locked pages…but TF 835 is on?

Hello Old Friend...AWE

- On Windows Lock Pages in Memory is Address Window Extensions (AWE)
  - Allocates contiguous mappings to PFNs. Logically contiguous, but not guaranteed contiguous
  - Linux will try to make the THPs contiguous
  - Then those PFNs are mapping into the process' virtual address space
- Why use AWE?
  - How are they unpageable?

Memory - What to look for?

- High consumers of space
  - Physical
  - Virtual
- External memory pressure on SQL Server
- Excessive swapping
  - swapping in/out

Tools to use for memory monitoring

- `/proc/meminfo`
- `free`
- `top/htop`
- `ps`
- `vmstat`
- `pidstat`

Demos

- Memory layout
- Isolating a memory hog
- Identifying external memory pressure
  - External memory pressure on SQL Server
- Excessive swapping
  - Swapping in/Swapping out

Disks and File Systems

Disks

- Sectors (physical)
  - Actual storage unit of the disk, 512B or 4KB
- Blocks (logical)
  - Fundamental unit of I/O, allocation
- Disks have finite performance characteristics
  - Bandwidth - how much data
  - Latency - how fast
- Storage Interconnects
  - Internal
  - External

File Systems

- XFS
  - Default file system - http://red.ht/2dBXccx
- EXT4
- Block size
  - Impact utilization and performance nominally
  - 4KB default block size
- Mount time options
  - Access times - `noatime`

Block Allocation in Linux

- XFS and EXT4 essentially the same
  - Files
  - i-nodes
  - Extents
    - Blocks

I/O under SQLPAL

- Stream I/O via NTUM
- Fast I/O via the host extension
  - Kernel asynchronous IO (kaio)
    - `io_submit()`
      - Returns to caller immediately, completion polling is in user space
  - `O_DIRECT` – bypasses page cache and I/O stays in user mode
    - `fsync()`
    - “probably designed by a deranged monkey on some serious mind-controlling substances.” - Linus
      - `man 2 open`

Disks - What to look for?

- **This is the slowest thing in your computer, sorry Argenis! :)**

- Saturated disks and I/O subsystems
- Swapping
- Caching is your friend (generally, but not in an RDBMS)
- Baseline!

Tools to use for disk monitoring

- `iostat`
- `iotop`
- `pidstat`
- `dstat`

Demos

- Finding high I/O processes
- Measuring disk latency (DMVs and cmd line tools)
    - `sys.dm_io_virtual_file_stats`

Monitoring Tools

Baselining Tools

- Nearly everything we've talked about so far has been point in time…what about baselining?
  - `sar` - System Activity Reporter
  - `dstat` - writes to CSV

Tools for Monitoring SQL Server

- You have all of the same tools you’re used to for SQL Server
  - Because of SQLOS we get
    - DMVs
    - Extended Events

New Tools Available for SQL on Linux

- New DMVs

- PSSDiag
  - https://blogs.msdn.microsoft.com/sqlcat/2017/08/11/collecting-performance-data-with-pssdiag-for-sql-server-on-linux/
- DBFS
  - https://github.com/Microsoft/dbfs
  - http://www.centinosystems.com/blog/sql/dbfs-command-line-access-to-sql-server-dmvs/
- Grafana
  - https://blogs.msdn.microsoft.com/sqlcat/2017/07/03/how-the-sqlcat-customer-lab-is-monitoring-sql-on-linux/

Metrics Captured by PSSDiag

- Don’t just listen to me…here’s what Microsoft is interested in
  - CPU - `mpstat`, `pidstat`
  - Disk - `iostat`, `iotop`
  - Memory - `free`, `sar`
  - Network - `sar`
  - DMV Data
  - System log information

Review

- Linux System Architecture
- SQL on Linux Architecture
- System Components
    - CPU/Processes
    - Memory/Pages
    - Disk/File Systems
- Monitoring Tools

Need more data?

Blog

www.centinosystems.com/blog

Pluralsight

Understanding and Using Essential Tools for Enterprise Linux 7

Linux basics, system architecture, file and directory management

LFCE: Advanced Network and System Administration

systemd, Performance and Tools

References

Many of the man pages

https://docs.microsoft.com/en-us/sql/linux/sql-server-linux-performance-best-practices

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/performance_tuning_guide/index

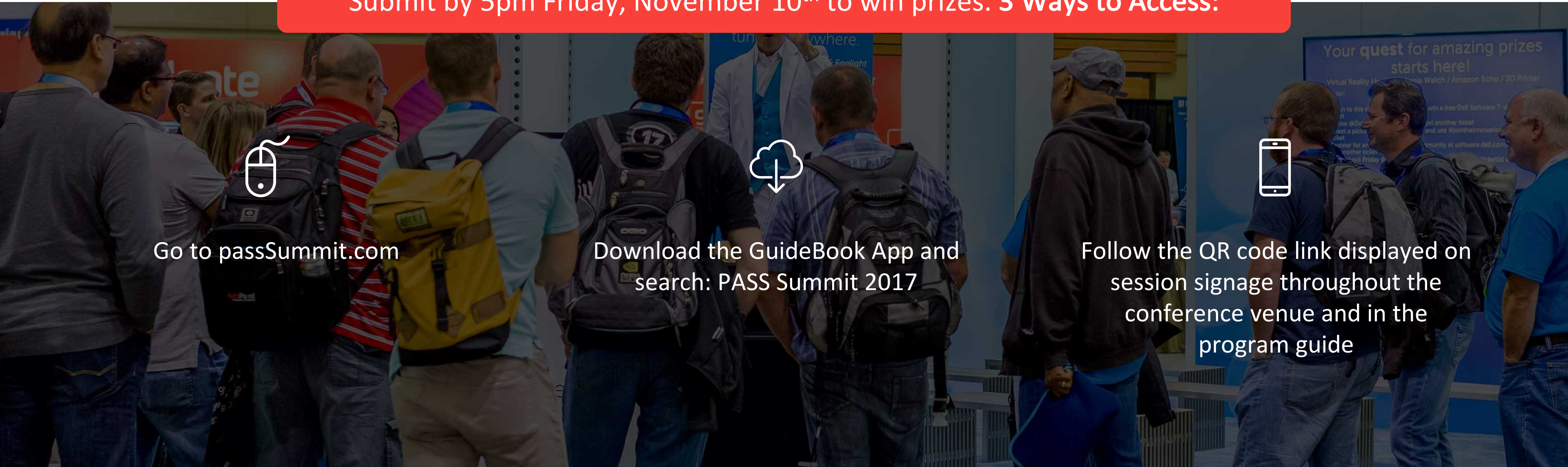
https://www.kernel.org/doc/Documentation/

https://ext4.wiki.kernel.org/index.php/Clarifying_Direct_IO%27s_Semantics

Session evaluations

Your feedback is important and valuable.

Submit by 5pm Friday, November 10th to win prizes. **3 Ways to Access:**



Go to passSummit.com

Download the GuideBook App and search: PASS Summit 2017

Follow the QR code link displayed on session signage throughout the conference venue and in the program guide

Thank You

Join me for the BOF lunch from 12P-2P

@nocentino

aen@centinosystems.com

PASS