Readout Server PRR status

Roland Sipos - for the DUNE DAQ
CERN

DUNE DAQ General Meeting
4th September 2024

Context

- Procurement of Readout Units for VD installation at SURF by November 2026
  - HD installation expected in mid 2027
- Current timeline: start of procurement at Q1 2025
  - Might be pushed for later dates
- Requires a Procurement Readiness Review (PRR) for the collaboration, that includes:
  - Procurement strategy
  - Technical Specification for tendering
- Multiple funding agencies are involved, which differ in internal procurement procedures and processes
  - With their custom constraints and deadlines

Deliverables

- Key items for the PRR readiness:
  - Documentation of funding agencies' strategies
  - Technical Specification for invitation to tender
  - Items that drive the technical specification decision

- Technical Specification includes:
  - Key items of the Specification of Technical deliverables:
    - System Units and Enclosures (next slide)
    - Common requirements (OS, BIOS settings)
    - Performance requirements (Computing, Storage, and Power draw)
  - Specification of the activities like packaging and shipping
  - Applicable rules, norms, and standards
  - Contract performance (Requirements on supply, acceptance, and warranty)

Servers' technical specification

- Processors
  - Cores, frequency and features
- Interconnects
  - Sockets, PCIe lanes, etc.
- Memory
  - Capacity, bandwidth and channels
- Storage
  - Type, capacity and bandwidth
- Network
  - Bandwidth and features (e.g.: RDMA)
- Chipset & mainboard
  - Features (e.g.: on-board accelerators)

	Minimum Requirements	Recommended Requirements
Drive type	HDD	SSD
CPU	4 cores (8 logical threads), frequency - 3-3.5 GHz and more	8 cores (16 logical threads), frequency - 3.5 GHz or more
RAM	8 GB or more	32 GB or more
Free disk space	200 GB or more	500 GB or more
Network interface bandwidth	100 Mbps	1 Gbps
HDD for IIS and documents	64 Gb	128 Gb
SSD for SQL	200 Gb	500 Gb

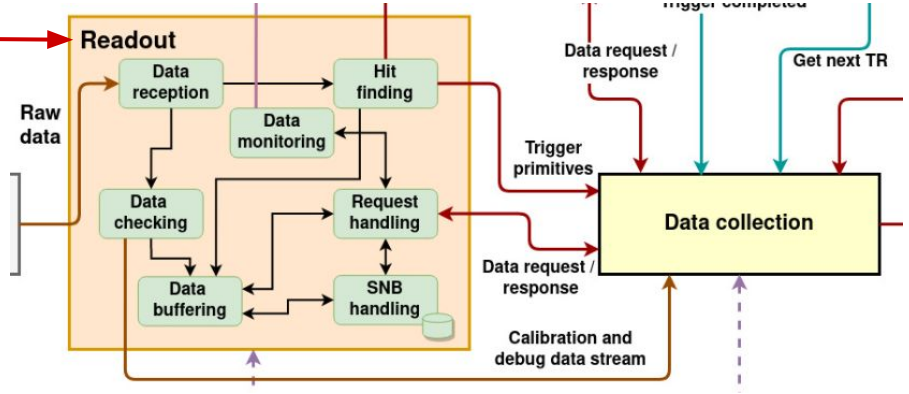
Readout system and RU specs

Several months of work was done (and ongoing) in order to provide the Technical Specification for the Readout servers

TABLE V
CHARACTERISTICS OF THE SCALE-UP DEMONSTRATOR SERVER

Component	Specification
Baseboard	Supermicro® X13DEM
CPU	Intel® Xeon® Gold 6448H @ 2.40 GHz (4.10 GHz turbo), 32-core 2S (dual socket) Code name: Sapphire Rapid 3 MiB L1d, 2 MiB L1i 128 MiB L2 120 MiB L3
DRAM	DDR5 1.0 TB, 4800 MT/s
NIC	2 x Intel E810-CQDA2 (1 per socket)
Drives	6 x 7.68 TB U.3 NVMe drives Samsung 980 Pro (3 per socket)
OS, DPDK	Alma Linux 9.3, Linux kernel 5.4, DPDK 22.11

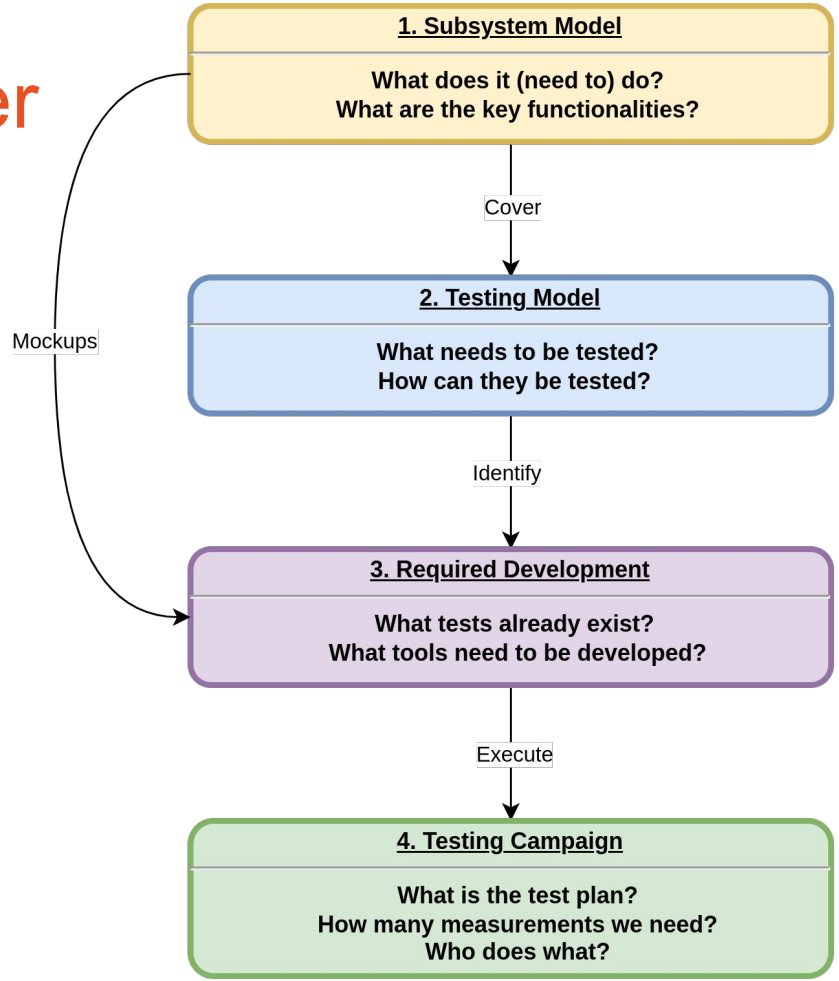
used for



Approach diagram reminder

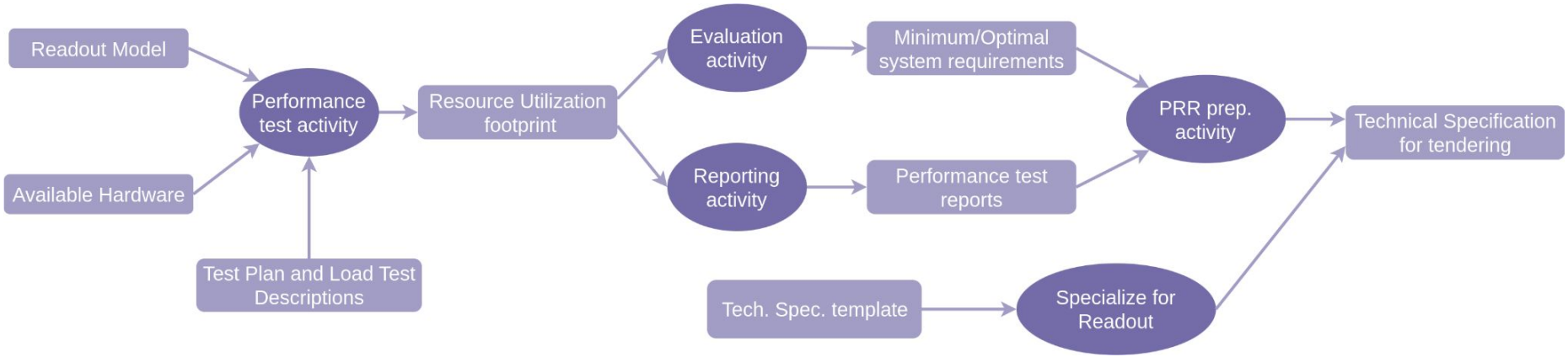
The following steps are needed to assess the technical specification of use-case servers:

1. Model of subsystems
2. Model for testing
3. Developments required for testing
4. Testing campaign



Readout Unit's approach

Generic approach diagram specialized for Readout Unit approach



Sub-deliverables

- Test plan descriptions
  - Scope: performance of readout system
  - Aim: gather resource utilization footprint of system in scope
  - System and software architecture description
  - List and description of system and component tests

- Performance test reports
  - Specifications of used hardware
  - Resource utilization footprint of each component
  - Component placement strategy (including isolation and affinity details)

- Clearly specified viable high-level configurations
  - For target throughputs of 200 vs. 400 Gbps
  - Symmetric or asymmetric component placement topologies

2 x FELIX hosts excluded!

Available hardware

Readout units located at EHN1 for NP04, NP02, and evaluation

SNB capture capable

- Currently in operations (excl. SNB store) for NP04:
  - Intel Skylake (launch 2016) np04-srv-021/022
  - Intel Cascade Lake (launch 2019) np04-srv-028/029

- For NP02:
  - Intel Ice Lake (launch 2020) np02-srv-002
  - AMD EPYC Zen3 (launch 2021) np02-srv-001

w/o drives

- For testing:
  - Intel Ice Lake - np02-srv-004 (from Canada)
  - AMD EPYC Zen3 - np02-srv-003 (from Canada)
  - Intel Sapphire Rapid (launch 2021) - np04-srv-031 (from CERN)

Performance test description

- Description of testing procedure for evaluation the servers' performance and specs.

- Similar to system tests, but with more emphasis on performance critical characteristics
  - Which subcomponent is assigned to what resources (Socket, PCIe, CPU, RAM)
  - Isolation and affinity techniques in place
  - Host and operating system optimizations in place

- Based on load test templates, specialized for readout. A typical structure:
  - Scope and approach
  - Systems under test environment
  - Performance and capability goals (aka.: pass/fail scenarios)
  - Load descriptions: testing process, tools used, status reporting
  - Test deliverables (e.g.: resources utilization, final report)

Test description example

- We didn't manage so far to establish proper test descriptions, we need to work on this more
- Aim for simplicity, but without hindering its usefulness
- Plan: Use existing Load Test templates
  - Keep important parts (e.g.: KPI)
  - Expand it with Readout specific items

Performance Test Plan Template	Revision 6/8/2021
Purpose:	
Definitions:	
Performance Criteria	
Goals:	
Test Type:	
Failure Criteria:	
Technical Requirements	
Environment:	
Credentials:	
Telemetry:	
Load Profile	
User Lifecycle:	
Concurrency:	
Post-Test Analysis	
Data collection:	
Reporting:	
Summary:	

Test report examples

- Used hardware
- Configuration
  - Topology and placement strategy
- Resource utilization during run for each thread:
  - CPU percentile
- Total system load
  - CPU
  - Caches
  - Memory bandwidth

TABLE III
SPECIFICATIONS OF THE INTEL INTEGRATION SERVER

Component	Specification
Baseboard	Intel® Server Board M50CYP2SBSTD
CPU	Intel® Xeon® Gold 6346 @ 3.10 GHz (3.60 GHz turbo), 16-core 2S (dual socket) Code name: Ice Lake 1.5 MiB L1d, 1 MiB L1i 40 MiB L2 72 MiB L3
DRAM	DDR4 512 GB, 3200 MT/s
NIC	Intel E810-CQDA2
OS, DPDK	Alma Linux 9.3, Linux kernel 5.4, DPDK 22.11

TABLE VI
OVERVIEW OF COMPONENTS AND THEIR RESOURCE NEEDS FOR 2 CRPS

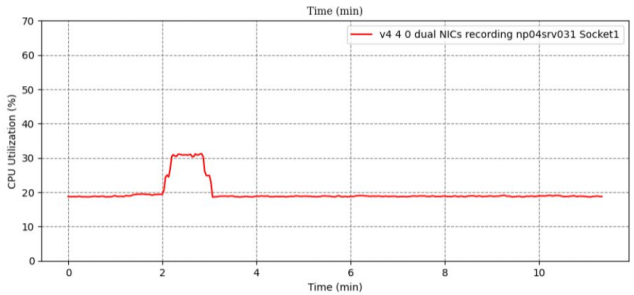
Component	Number of threads	CPU cores assigned	Maximum CPU core utilization (%)
Data reception (Packet processors)	8	4 phys. and 4 HT [a]	~48.2
Data processing (TPG)	96	10 phys. and 10 HT	~55.8
Supernova Burst (Recording)	96	8 phys. and 8 HT	~52.6

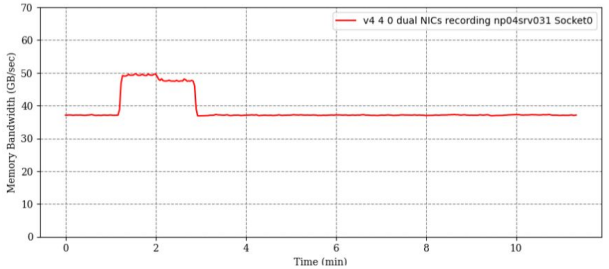
[a]CPU cores are assigned with their corresponding Hyper-thread (HT) core included.

Configurations:

* cpupin-eth-mockdlh_grouped-np02srv004-1.json

Pinning	CPU cores
parent	64-127,192-255
cleanup-.	64-71,192-199
producer-.	72-79,200-207
consumer-.	80-87,208-215





System coverage matrix

1. List of components mapped to server spec. needs

Component	Devices and interconnects	CPU	Memory	Persistent storage
Data reception	NICs and PCIe lanes	sensitive	sensitive	
Latency Buffer	Memory and its channels	marginal	sensitive	marginal
Data processing	CPU and cache lines	sensitive	sensitive	
Supernova Burst Data Store	Persistent storage	marginal	sensitive	sensitive

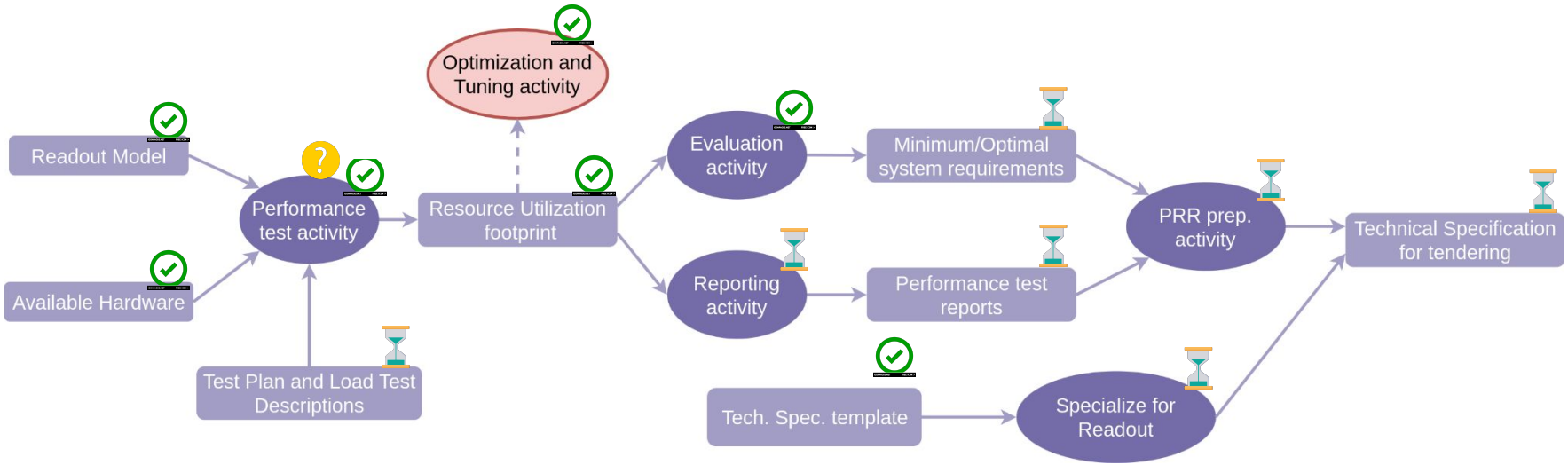
2. Methods for quantifying the resources needs

3. Tools to be developed tied to a list of methods
   Green: Developed
   Red: Refinement needed

4. Test plan for measurements tied to tools and methods
   OK or ongoing: 

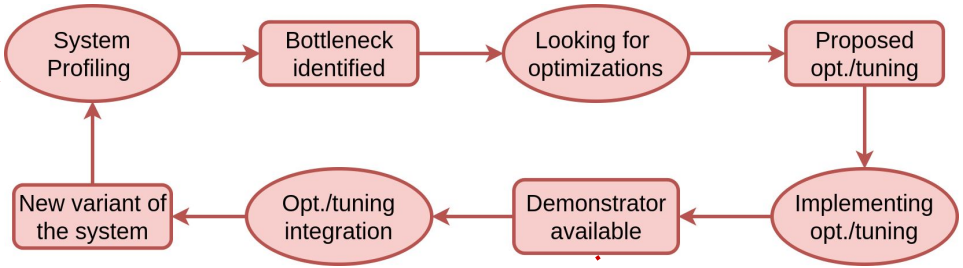
Components/ what needs to be tested	Device feature and interconnects	CPU utilization	Memory utilization	Storage utilization
Data reception	Can be calculated, Acceptance tests of 100Gb NICs	Test 1.: DPDK reception Test 2.: Copy vs. callback Test 3.: integrated system no missed/dropped packets	Can be calculated, cross-checked with PCM (~10GB/s per 100G)	
Latency Buffer	Can be calculated, Max bandwidth I/O	Test 1.: Prod/consumer/ request rate stress tests	Can be calculated, Test 1.: maximum throughput tests	Tests: filewriter and LB to drive via zero copy
Data Processing	Cache size and locality sensitive, AVX2 capable CPU	Tests: TPG algo., emulator tests, TPG rate scaling, Integrated system (A. Oranday)		
SNB capture	Can be calculated, High-speed NVMe	Can be calculated, cross-checked with standalone benchmarks	Can be calculated, cross-checked with standalone benchmarks	Can be calculated, cross-checked with standalone benchmarks

Approach status

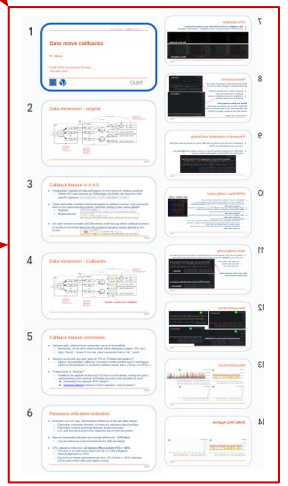


Optimization and tuning activity

- Steps of a single stage of the optimization pipeline looks like:



- Few examples (stages) of this work:
  - Component placement differences
  - AMD DMA latency
  - SNB store (RAID0, off-kernel, zero-copy mode)
  - Kernel isolation of data reception
  - Data insert callbacks for Latency Buffers
  - 100Gb NIC optimal polling config via DPDK
  - CPU sleep states, AVX mixed workloads, etc.



Milestones

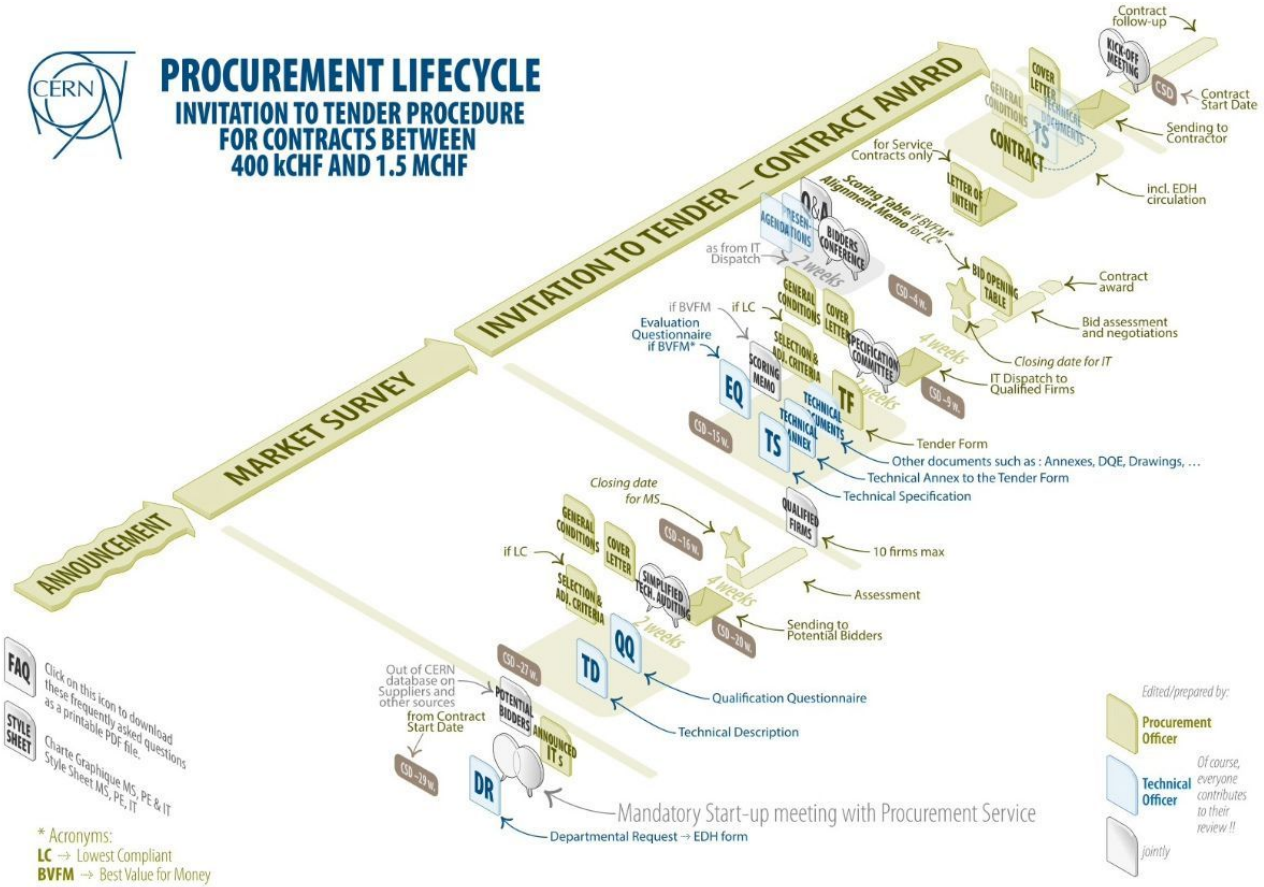
- Performance test activity completed by Q4 2024 for currently available hardware. (Result: resource utilization footprint)
- Reporting activity completed by Q1 2025 (Result: Performance Test Report documents available on EDMS)
- Evaluation activity completed by Q2 2025 (Result: Minimum/Optimal system requirements)
- Readout specialized Technical Specification for Tendering ready by Q4 2025
  - PRR right after Tech. Spec. is ready
  - Launch of procurement right after PRR (essentially ~1 year before delivery to SURF)

Standpoints on current strategy

- When is the latest date for baseline change from 200Gb to 400Gb aggregation per Readout Unit?
  - Essentially now or as soon as possible (end of September?), as it has many implications (e.g.: downstream network) and reduces testing needs

- Prioritization: we should focus on running all remaining tests, focused on extracting the necessary information, and document them in form of reports

- Clarify possible extra constraints of procurement strategies
  - Departmental request dates and approvals differ among funding agencies (Canada, UK, CERN)
  - Major planning overhead for common procurement



https://procurement-lifecycle.web.cern.ch/ProcLC-400k-1.5M.html

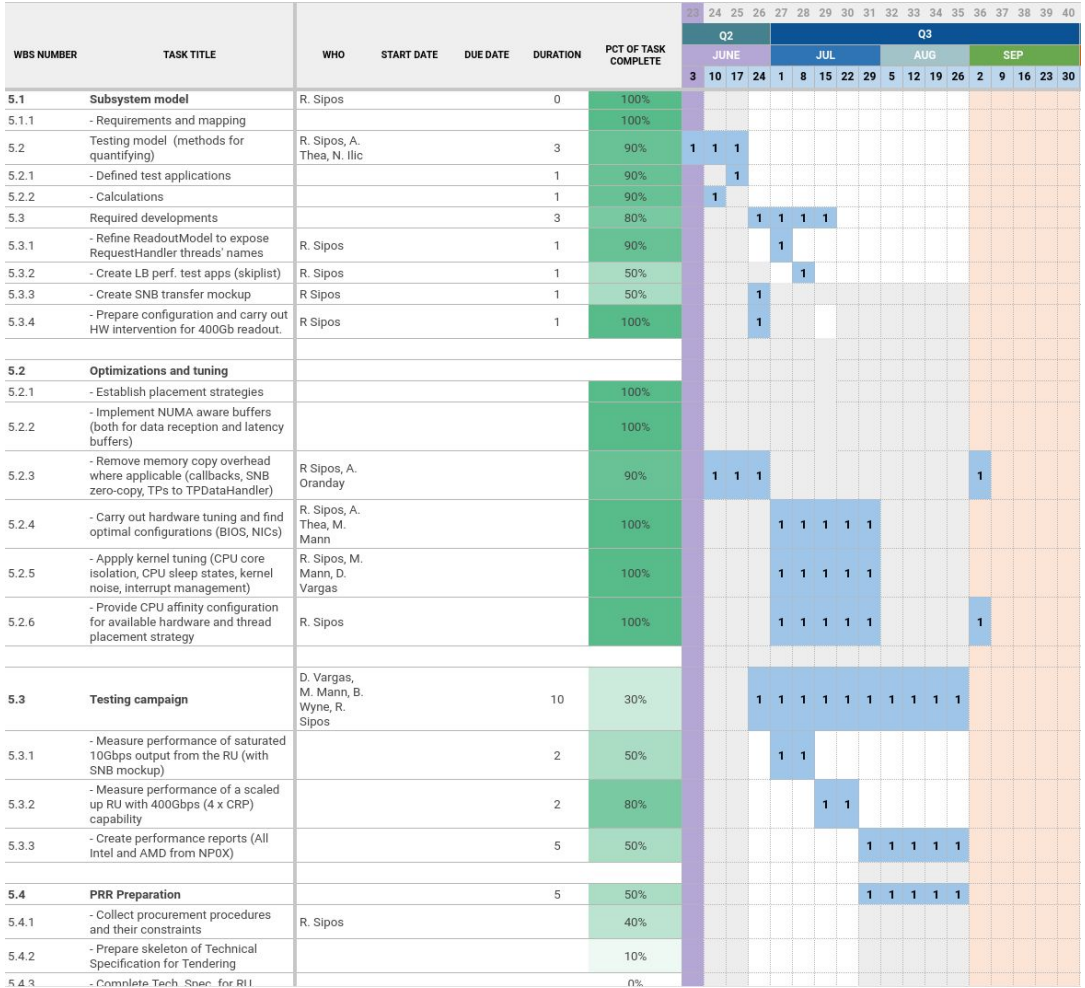
Imminent problems and risks

- Performance testing activity should be as automated as possible, which is not the case at the moment
  - New hardware should be able to be tested with as minimal effort as possible
- Test description activity is not complete, and reporting activity needs adjustments
- Personnel departures: 2 FTE just left
  - Very hard to plan timelines, where the effort essentially dropped to 0.1 FTE
- Keeping both 200Gb baseline and 400Gb proposed aggregation under the radar introduces major overhead in terms of planning and test execution

PRR Gantt

- [Link]
- Lot of effort invested in optimization and tuning of the readout system, that is not really a part of the PRR activity
- Difficult to maintain due to priority shifts and available effort

WBS NUMBER	TASK TITLE	WHO	START DATE	DUE DATE	DURATION	PCT OF TASK COMPLETE
5.1	**Subsystem model**	R. Sipos			0	100%
5.1.1	- Requirements and mapping					100%
5.2	Testing model (methods for quantifying)	R. Sipos, A. Thea, N. Ilic			3	90%
5.2.1	- Defined test applications				1	90%
5.2.2	- Calculations				1	90%
5.3	Required developments				3	80%
5.3.1	- Refine ReadoutModel to expose RequestHandler threads' names	R. Sipos			1	90%
5.3.2	- Create LB perf. test apps (skiplist)	R. Sipos			1	50%
5.3.3	- Create SNB transfer mockup	R Sipos			1	50%
5.3.4	- Prepare configuration and carry out HW intervention for 400Gb readout.	R Sipos			1	100%
5.2	**Optimizations and tuning**					
5.2.1	- Establish placement strategies					100%
5.2.2	- Implement NUMA aware buffers (both for data reception and latency buffers)					100%
5.2.3	- Remove memory copy overhead where applicable (callbacks, SNB zero-copy, TPs to TPDataHandler)	R Sipos, A. Oranday				90%
5.2.4	- Carry out hardware tuning and find optimal configurations (BIOS, NICs)	R. Sipos, A. Thea, M. Mann				100%
5.2.5	- Appply kernel tuning (CPU core isolation, CPU sleep states, kernel noise, interrupt management)	R. Sipos, M. Mann, D. Vargas				100%
5.2.6	- Provide CPU affinity configuration for available hardware and thread placement strategy	R. Sipos				100%
5.3	**Testing campaign**	D. Vargas, M. Mann, B. Wyne, R. Sipos			10	30%
5.3.1	- Measure performance of saturated 10Gbps output from the RU (with SNB mockup)				2	50%
5.3.2	- Measure performance of a scaled up RU with 400Gbps (4 x CRP) capability				2	80%
5.3.3	- Create performance reports (All Intel and AMD from NP0X)				5	50%
5.4	**PRR Preparation**				5	50%
5.4.1	- Collect procurement procedures and their constraints	R. Sipos				40%
5.4.2	- Prepare skeleton of Technical Specification for Tendering					10%
5.4.3	- Complete Tech. Spec. for RU					0%



Summary

- The Readout System was optimized and fine-tuned thanks to the results of the performance testing activity
- Configuration and topology of the system that are affecting the Readout Unit resource utilization and performance is well understood
- Documenting the tests and reporting on their results needs more attention and work
- Revision of PRR strategy is needed in order to reduce effort on necessary test planning, execution, and also on procurement preparation activities
  - Baseline aggregation policy change (200Gb vs. 400Gb)
  - Funding agencies (may) have different constraints on Market Survey & Tendering