Architecture of ELT 1st light instruments' Hard Real Time Computing Facility *with Xeon-Phis*

Carlos M. Correia

- on behalf of -

Frédéric Chazalet, Didier Rabaud,
Léonardo Blanco, Thierry Fusco, Jean-François Sauvage,
Benoit Neichel, Noah Schwartz

Abstract

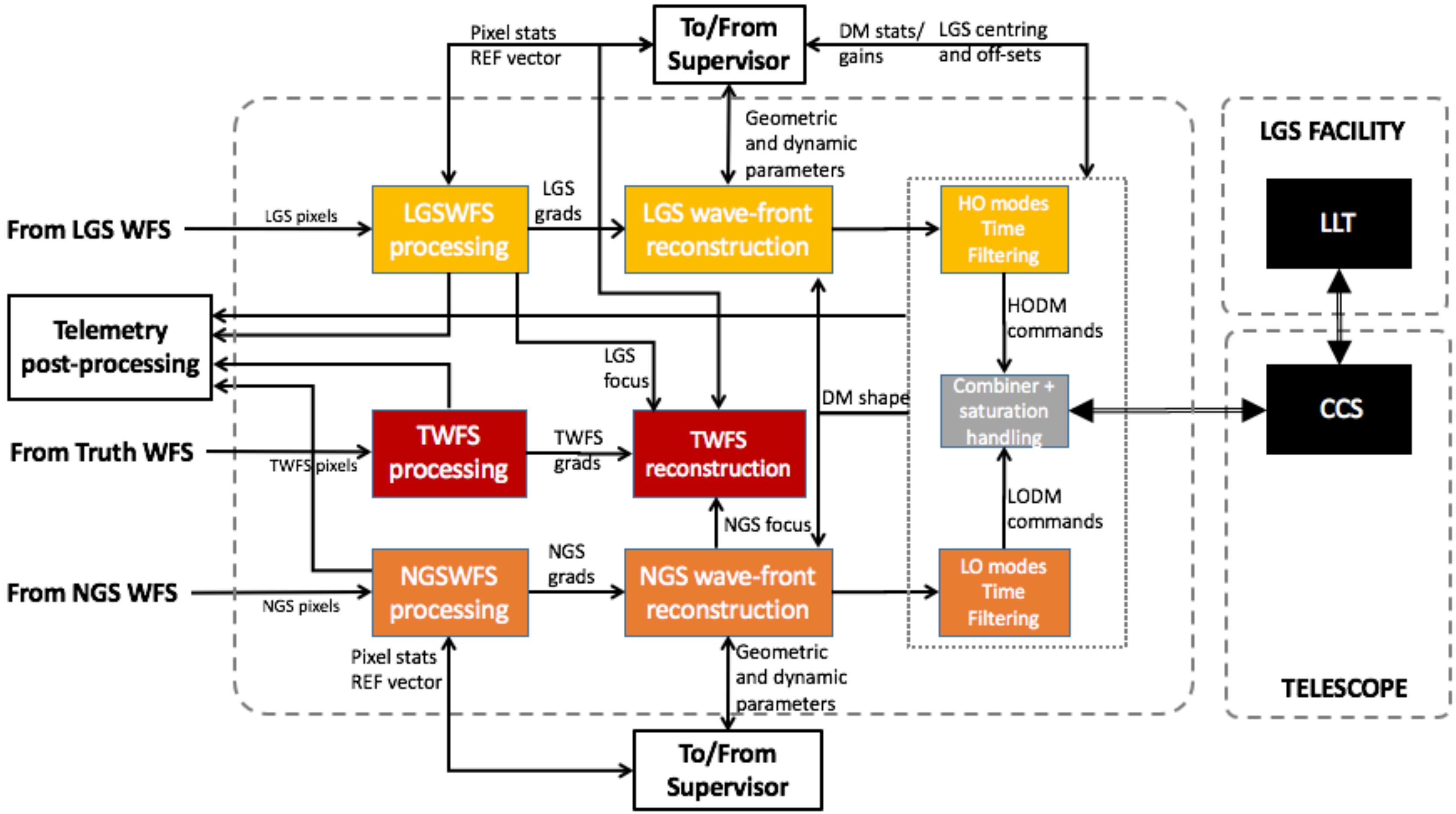
- We provide an overview of the key performance aspects covering computing power, memory bandwidth and throughput required for the Hard Real-Time Computer of the ELT 1st light instrument suite.
- An architecture is proposed for the most demanding tomographic wave-front reconstruction steps. Based on a detailed examination of the temporal diagram, backed by benchmarking results obtained with Xeon Phi processors, we find that with 1 node per Laser-Guide-Star Wave-front Sensor we can accomplish the Pixel Processing and Wave-front Tomographic Reconstruction steps (composed of wave-front reconstruction and Pseudo-Open-Loop slope calculation) within 1 loop cycle of 2ms with a latency that is below 1ms. We provide alternative arrangements should there be added overheads in any of the computation stages, granting the above figure is always met.
- We propose an extra node for the Time Filtering step and two other for the supervision and data recording. In total these amount to 9 if we tackle the full-scale problem.

Rationale

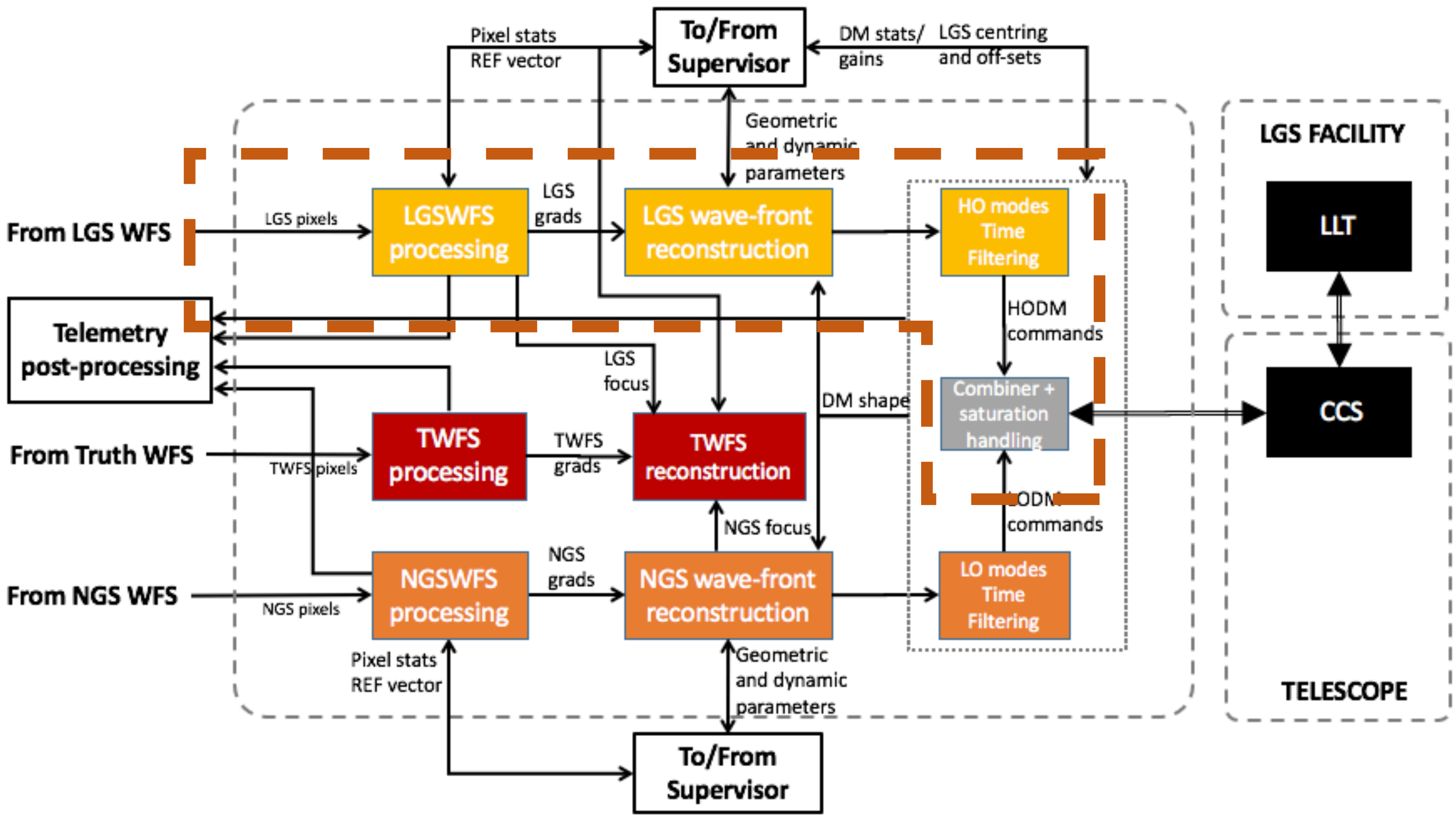
- Design and prototype a Hard Real Time Controller for Adaptive Optics
    - Tackle the tomographic cases on the ELT

- Technologies
    - x86-64 architecture, multi-core, multi-CPU, general purpose computers, including support for SIMD extensions and vector processing units
    - Many Integrated Core (MIC) architecture, **Intel Xeon Phi processors**

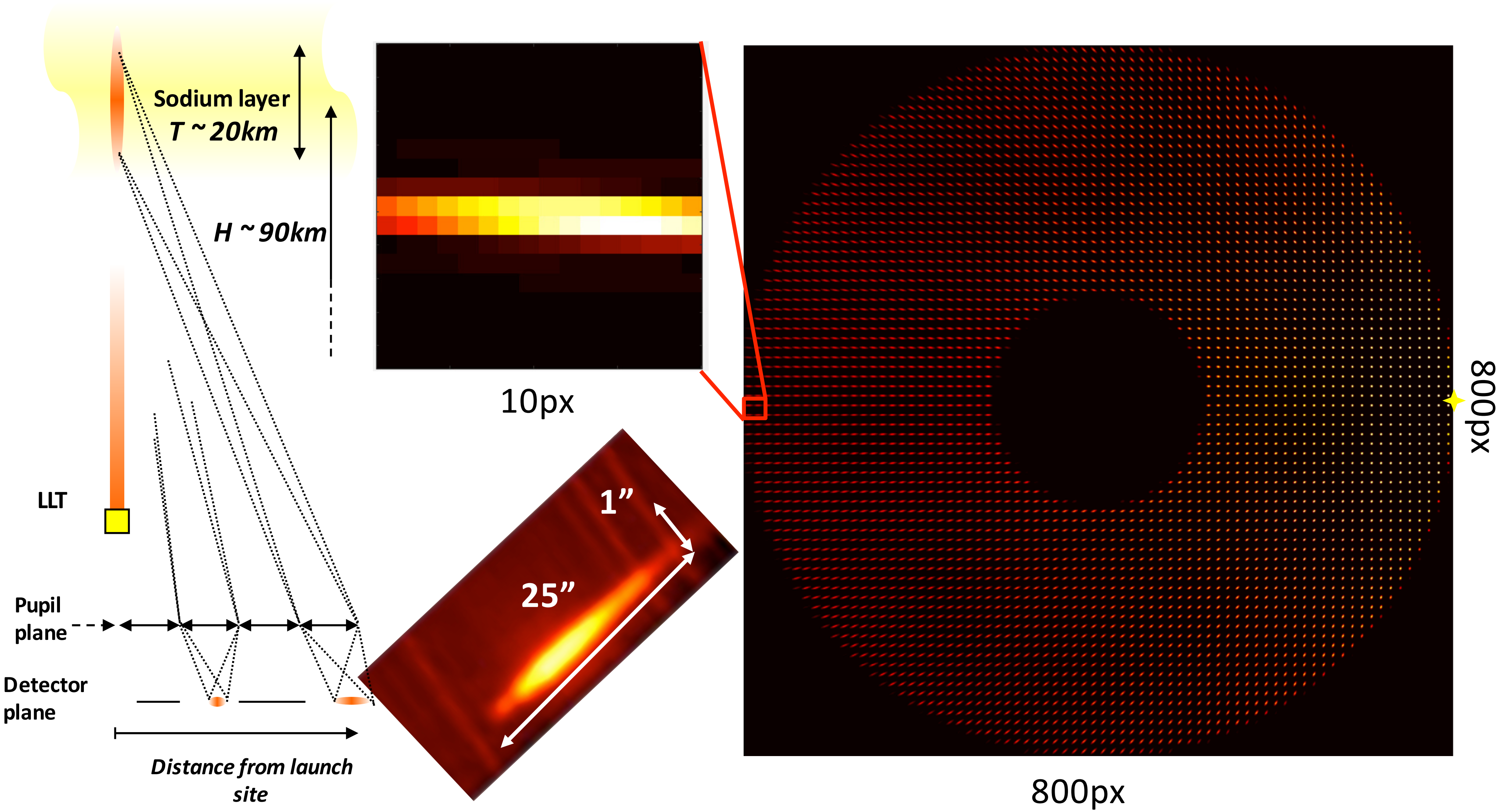
Conceptual Block Diagram



Conceptual Block Diagram



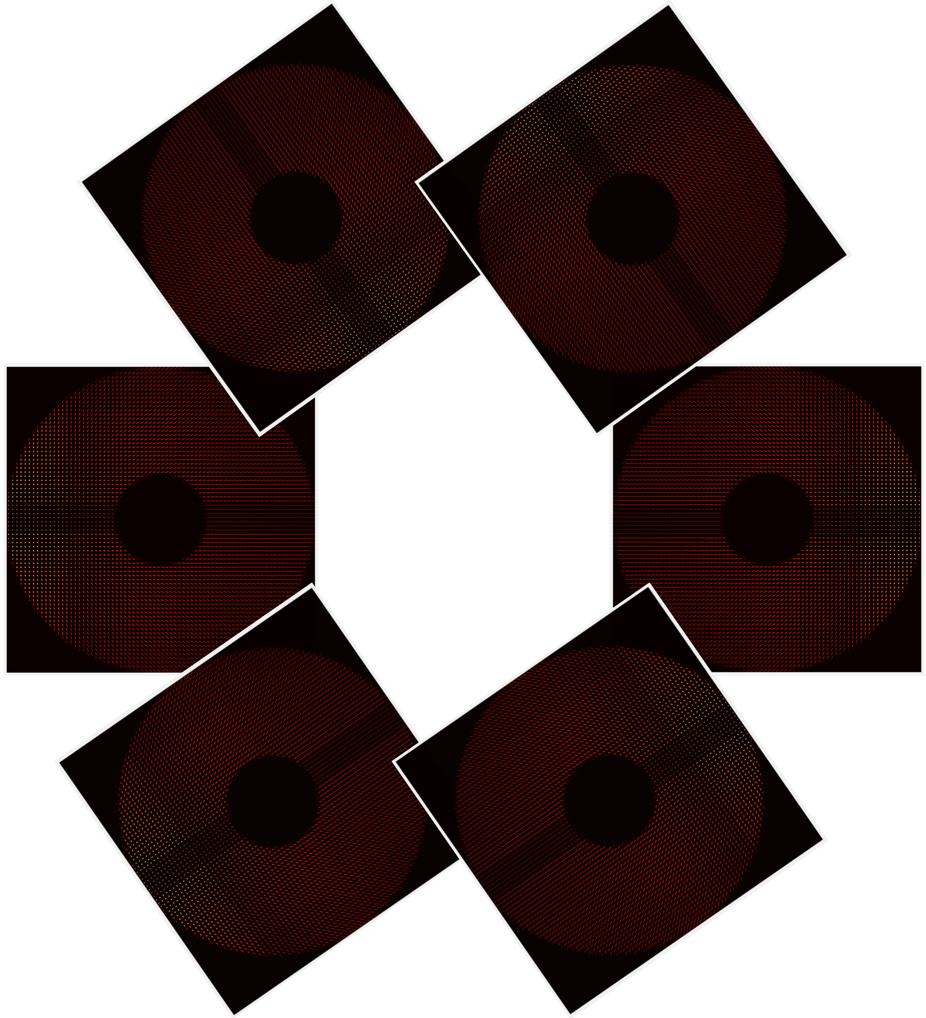
LGS WFSensing



Problem statement

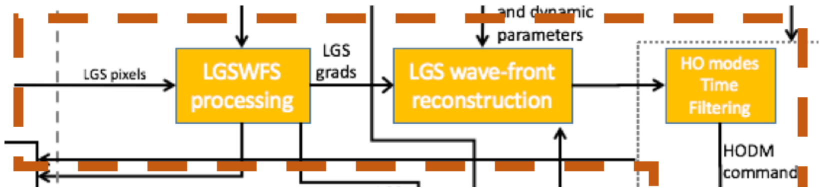
- Hard Real Time Controller for Tomographic Adaptive Optics
  - 6x 80x80 SH LGS WFS
  - 3x DM (M4: 5316, M2: 2, INS-DM:1000)

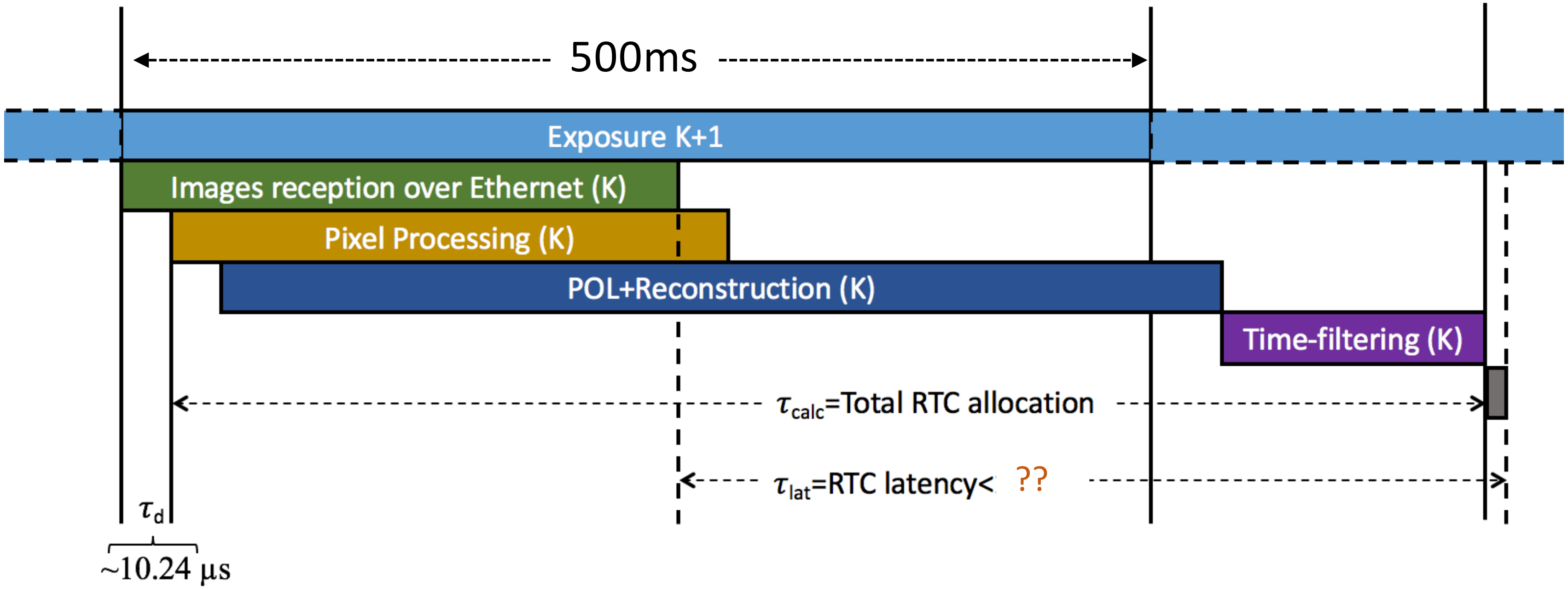
	Per LGS WFS
Total Number of Pixels	640000
Number of Slopes	9232
Commands	6316
Telemetry	Slopes + commands + (sub-sampled) detector pixels
Disturbance data	Slopes + commands
Matrices (M and R)	9232 x 6316



Operations

1. Pixel processing
2. POL + Reconstruction
3. Time-filtering

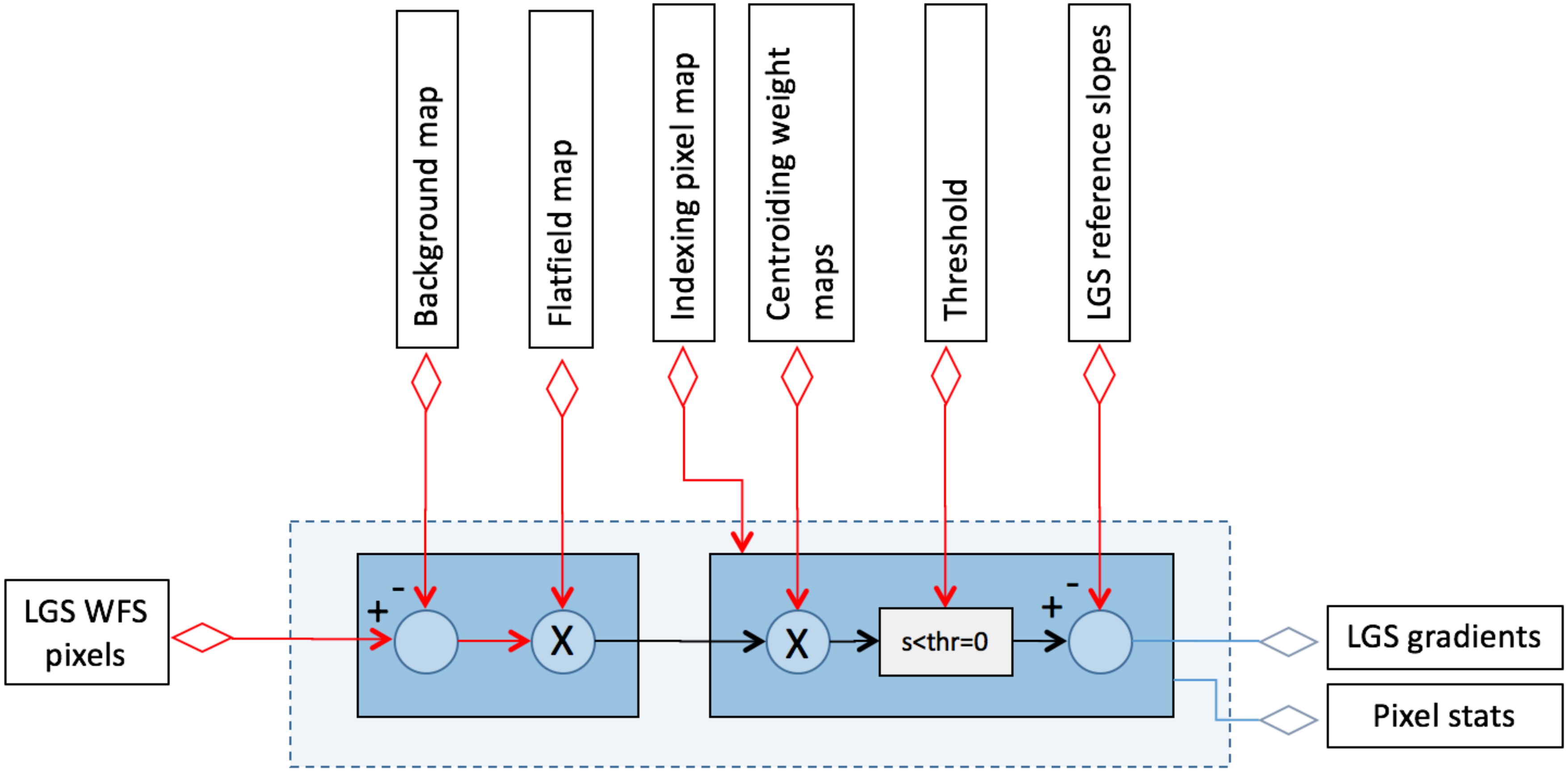




Reconstruction + POLC requirements

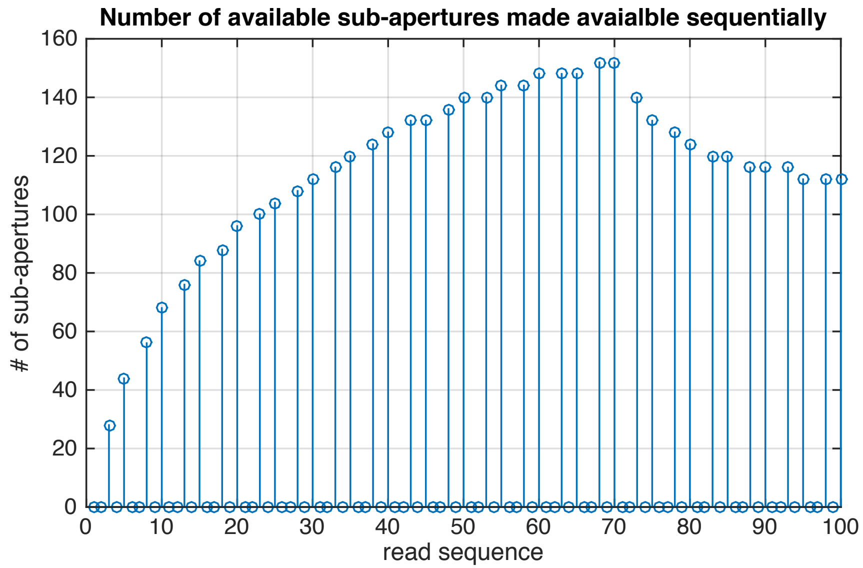
Task	Sub-task	# operations		Memory bandwidth			Memory storage
		# GFLOP/loop cycle	# GFLOP/s	Mbit/loop cycle	Gbit/s	Gbytes/s	Mbytes
WFS Processing	Calibration	1,28E-03	0,64	81,92	40,96	5,12	10,24
	Centroiding	3,21E-03	1,60	102,70	51,35	6,42	10,30
	Reference subtraction and Under-illumination handling	9,23E-06	4,62E-03	0,44	0,22	0,03	0,04
	Slope Disturbance Injection	9,23E-06	4,62E-03	0,44		0,00	0,04
WF Reconstruction	POL	3,79E-05	0,02				
		1,17E-01	58,31	1867,00	933,50	116,69	233,34
	Tomography	1,17E-01	58,31	1866,49	933,24	116,66	233,26
Temporal Filtering	Error estimation	6,32E-06	0,00	1,62	0,81	0,10	0,15
	Low-order removal	5,68E-05	0,03	1,62	0,81	0,10	0,18
	IIR	1,52E-04	0,08	2,63	1,31	0,16	0,30
	M4/M5 TT splitting	5,32E-05	0,03	1,02	0,51	0,06	0,03
	Command biasing	6,32E-06	3,16E-03	0,61	0,30	0,04	0,03
	Disturbance injection	6,32E-06	3,16E-03	0,61	0,30	0,04	0,03
	Total (1/6 of full problem size)	**0,24**	**119,03**	**3927,09**	**1963,32**	**245,42**	**487,92**

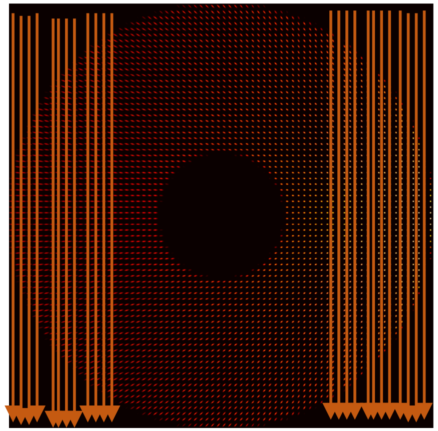
Pixel processing

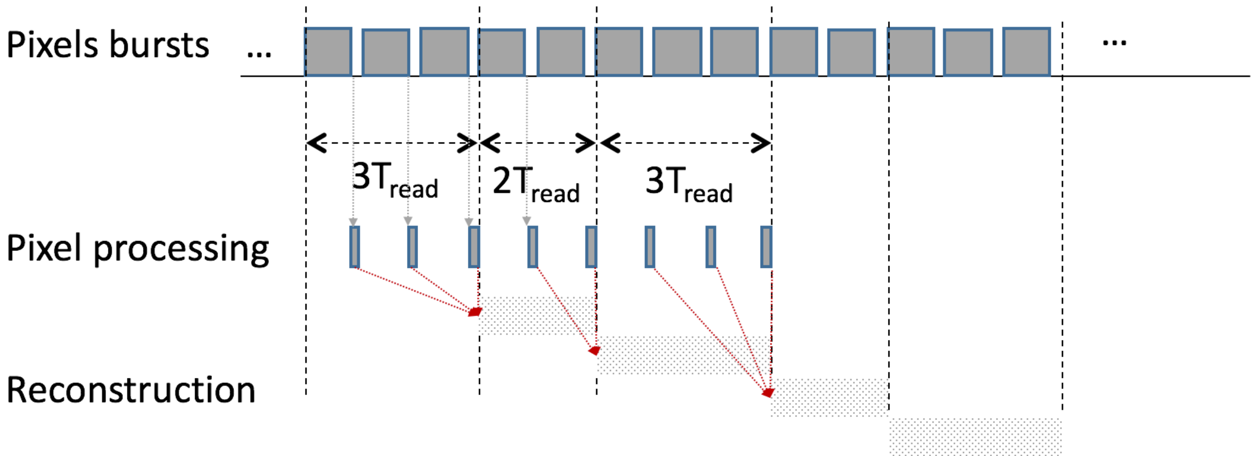


Pipelining the processing

- Current detector is read out in bursts of 8 lines at a time
  - (four from either side of the detector).
- 8x800x16bit = 102400bits sent over the Ethernet link in bursts.
- With 10GbE network links: 102400/1e10x1e6=10.24 μs+10-20% overhead 11.3-12.3 μs.

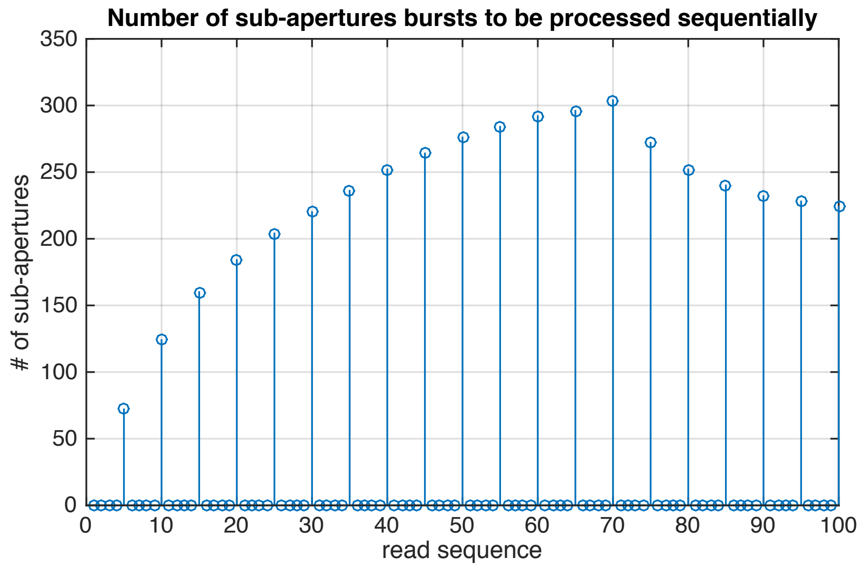


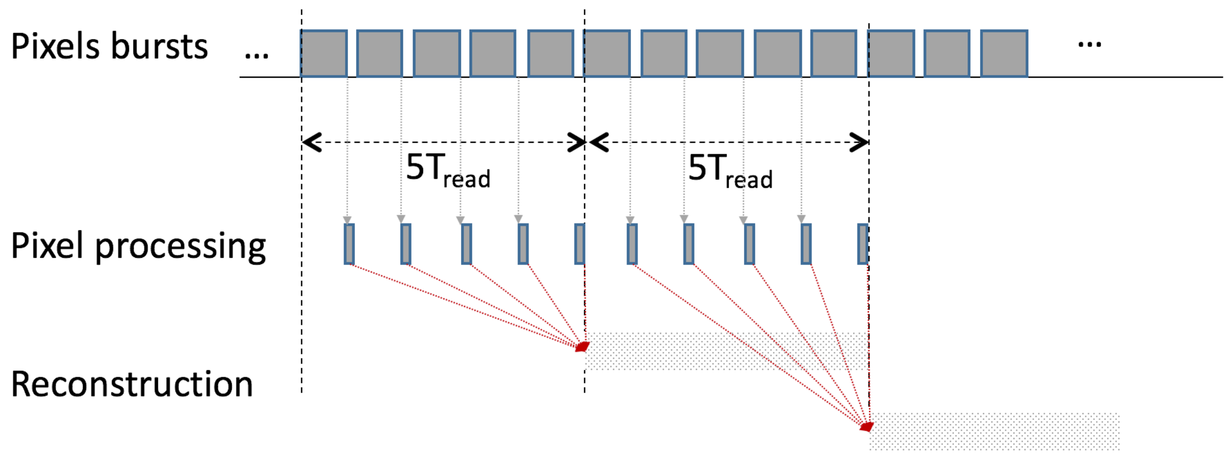




Pipelining the processing

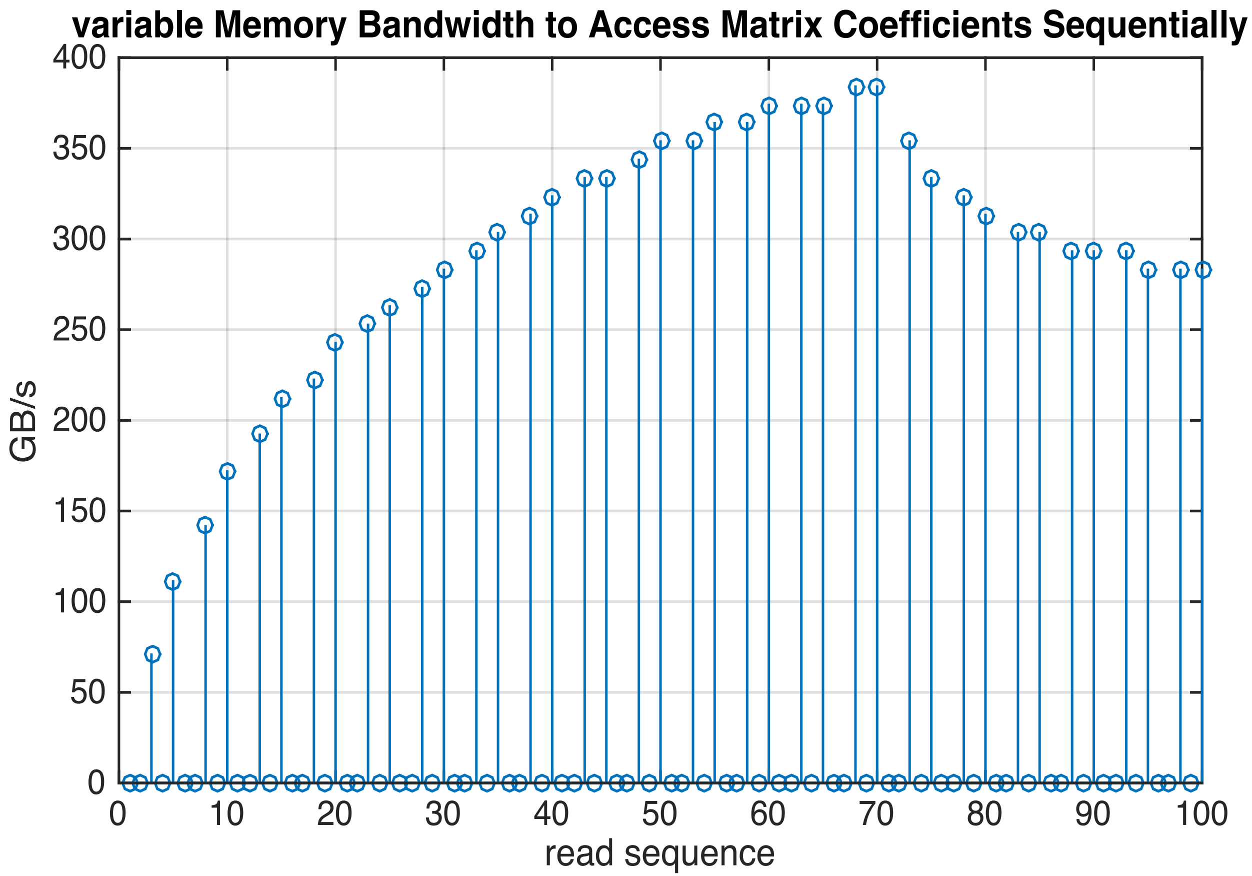
- Current detector is read out in bursts of 8 lines at a time
  - (four from either side of the detector).
- 8x800x16bit = 102400bits sent over the Ethernet link in bursts.
- With 10GbE network links: 102400/1e10x1e6=10.24 μs+10-20% overhead 11.3-12.3 μs.





Pipelining the processing

- Variable Bandwidth Memory Access Requirements



High-order MMSE tomographic reconstruction

- pseudo-open-loop measurements given by

$$s_k^{POL} = s_k + M\left(\sum_{i=1}^{2} \delta_i u_{k-1-i}^{CCS}\right)$$

where the scalars δ_i represent the relative contributions of commands integrated over the WFS sampling time such that

$$\delta_1 + \delta_2 = 1$$

- LO order mode removal

$$s_k^{POL,LOR} = (I - s2LO)s_k^{POL}$$

- Tomographic reconstruction

$$\widehat{\phi_\beta} = W s_k^{POL,LOR}$$

- Change of space to DM-commands space

$$u_\beta^{LGS} = F_{M4}\widehat{\phi_\beta}$$

High-order MMSE tomographic reconstruction and pseudo-open-loop control

- pseudo-open-loop measurements given by

$$s_k^{POL} = s_k + M\left(\sum_{i=1}^{2} \delta_i u_{k-1-i}^{CCS}\right)$$

where the scalars δ_i represent the relative contributions of commands integrated over the WFS sampling time such that

$$\delta_1 + \delta_2 = 1$$

- LO order mode removal

$$s_k^{POL,LOR} = (I - s2LO)s_k^{POL}$$

- Tomographic reconstruction

$$\widehat{\phi_\beta} = W s_k^{POL,LOR}$$

$$w_k = R(s_k + Mp_k)$$

- Change of space to DM-commands space

$$u_\beta^{LGS} = F_{M4}\widehat{\phi_\beta}$$

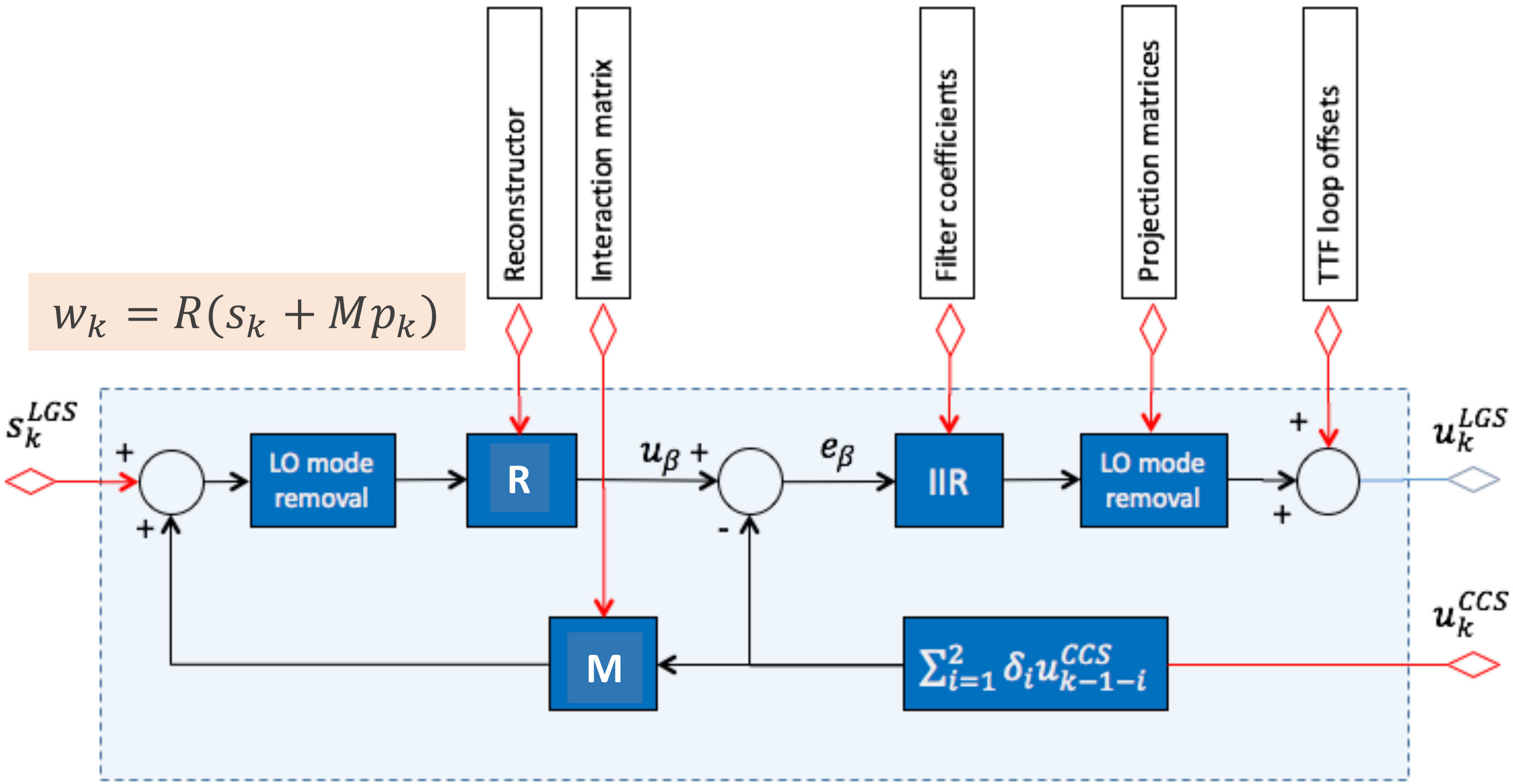
- Vector operation to remove mirror commands previously added in when performing pseudo-open-loop calculation (on step 1)

$$e_\beta = u_\beta^{LGS} - \sum_{i=1}^{2} \delta_i u_{k-1-i}^{CCS}$$

- Time filtering through IIR filter

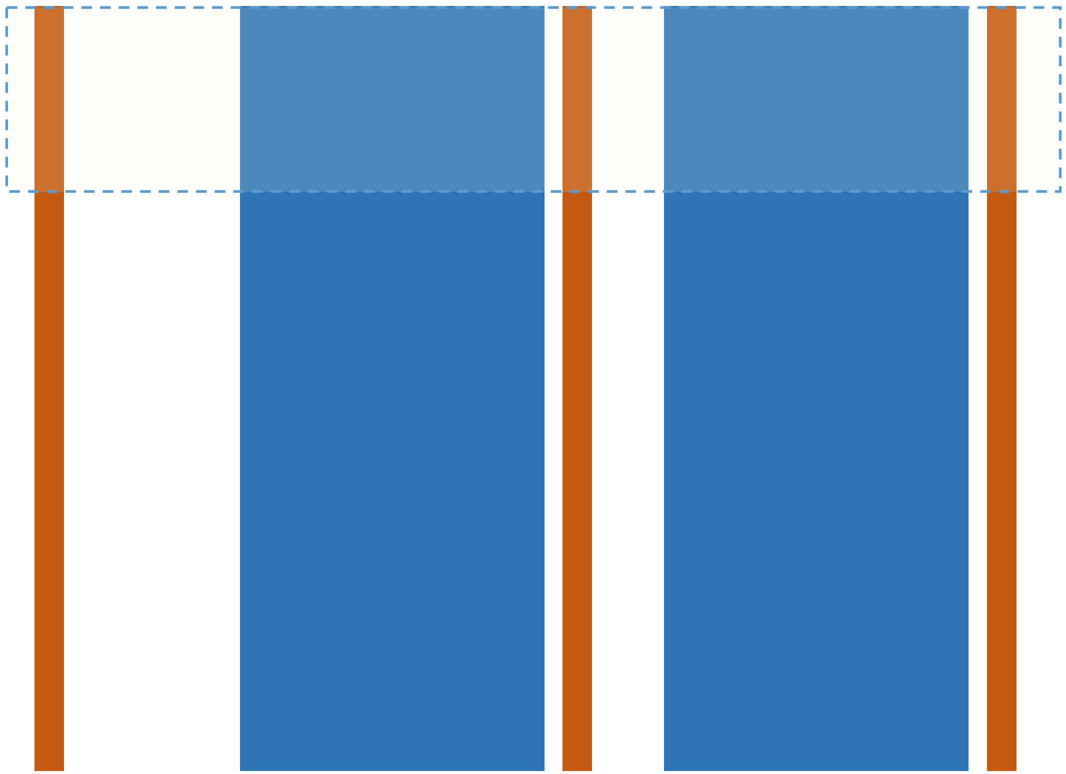
$$\mathrm{u}_k^{LGS} = \sum_{j=1}^{m_{IIF}} b_i \cdot (e_\beta)_j - \sum_{i=1}^{n_{IIF}} a_i \cdot u_{k-\mathrm{i}}^{LGS}$$

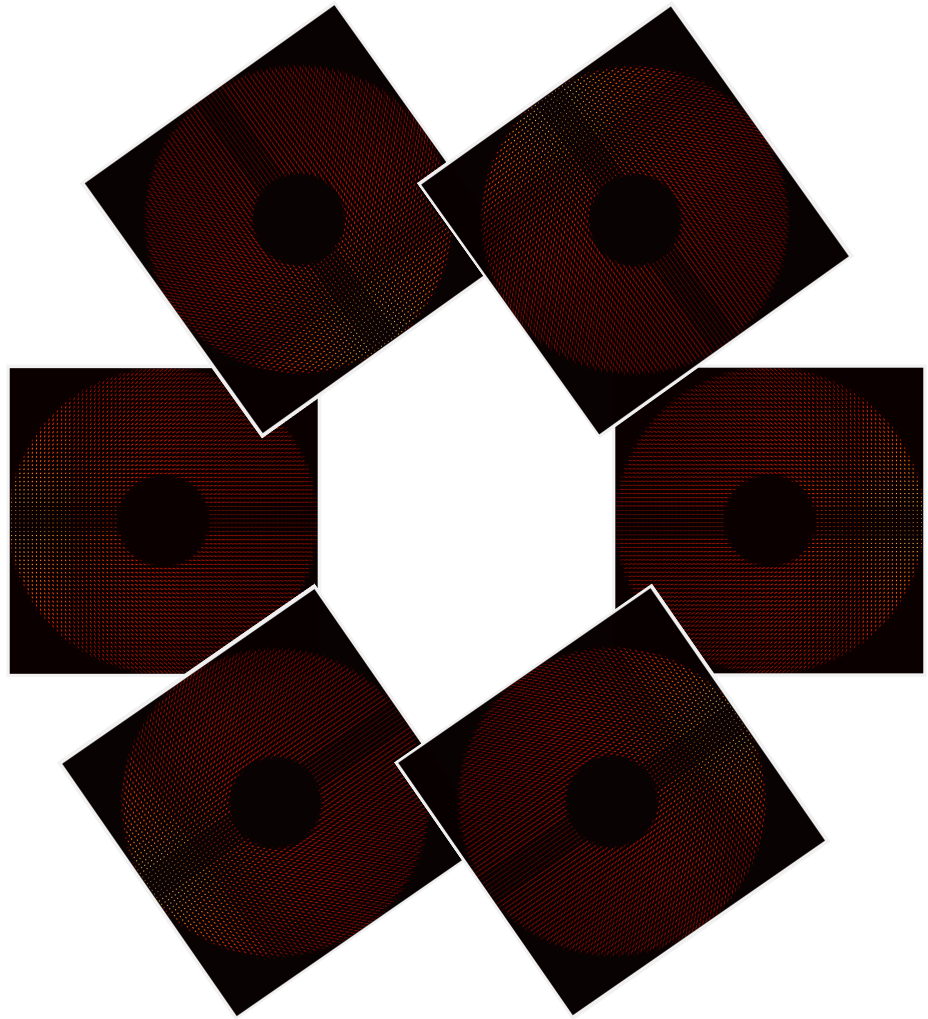
High-order MMSE tomographic reconstruction and pseudo-open-loop control



1/6th of the full problem

$$w_k = R \quad (s_k + M \quad p_k)$$





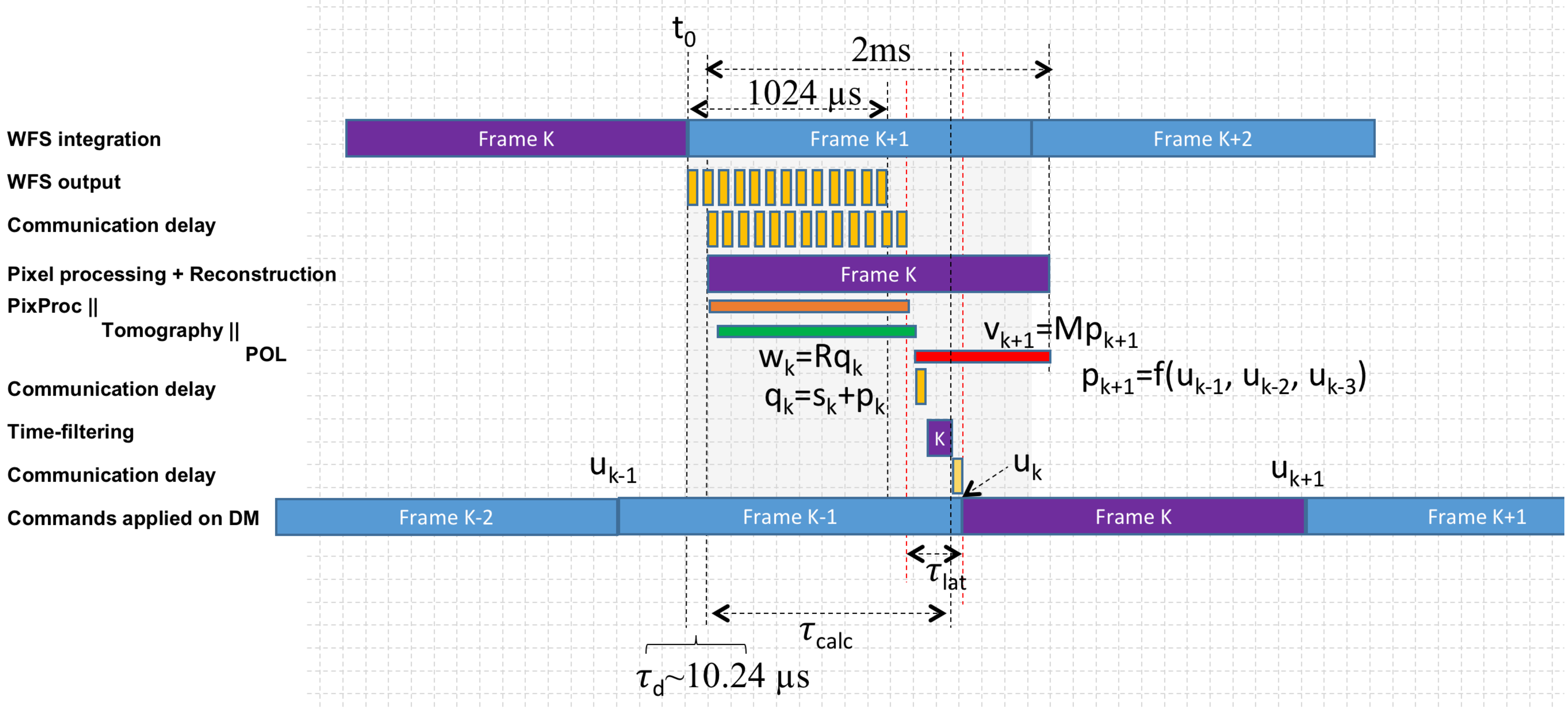
Initial benchmarking

- Pixel Processing followed by POL and Reconstruction, i.e. processing pixels to compute s_k, then compute Mp_k and Rq_k with $q_k = s_k + Mp_k$ takes 1870 μs.

- It motivates the temporal arrangement of operations laid out in the next section.

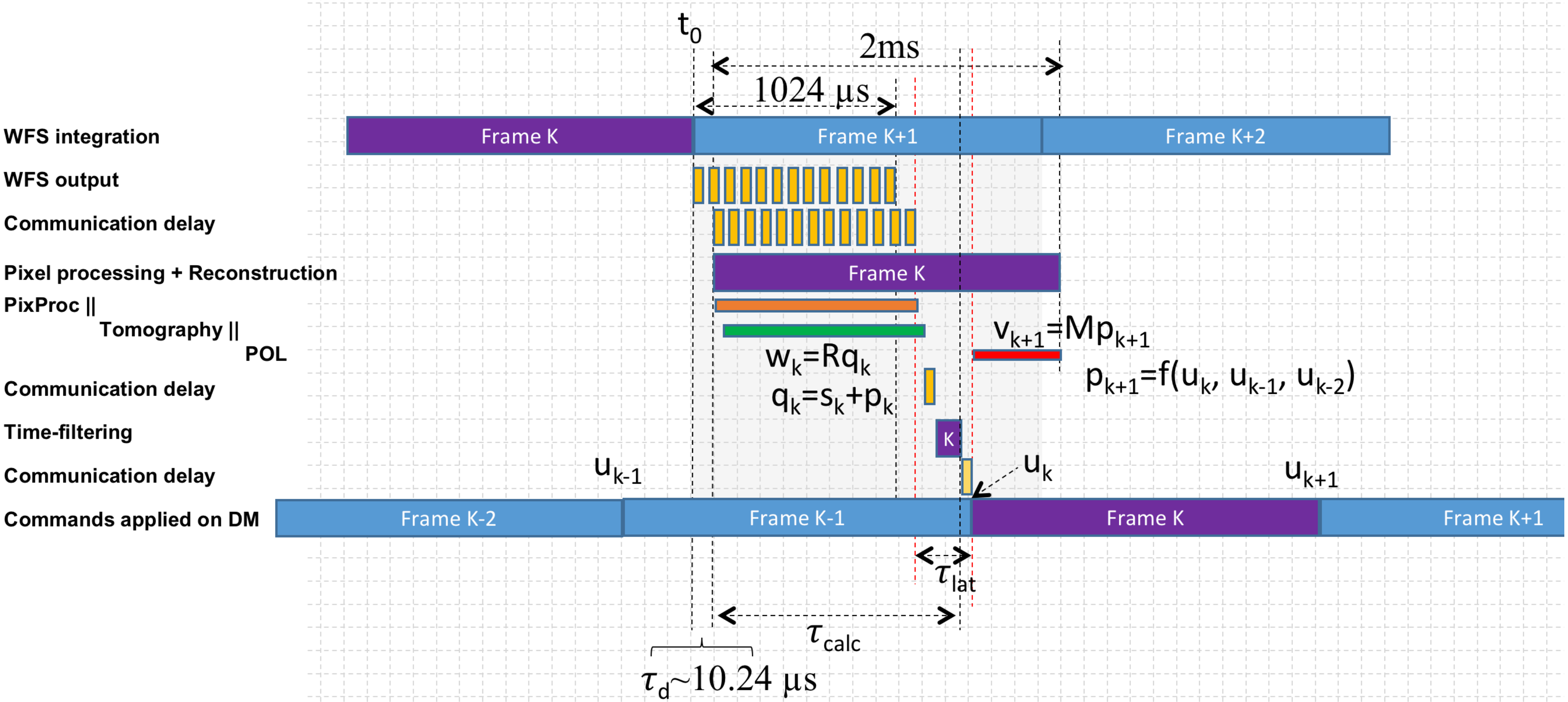
Time sequencing

- Case $\delta \geq 1$



Time sequencing

- Case $\delta = 0$



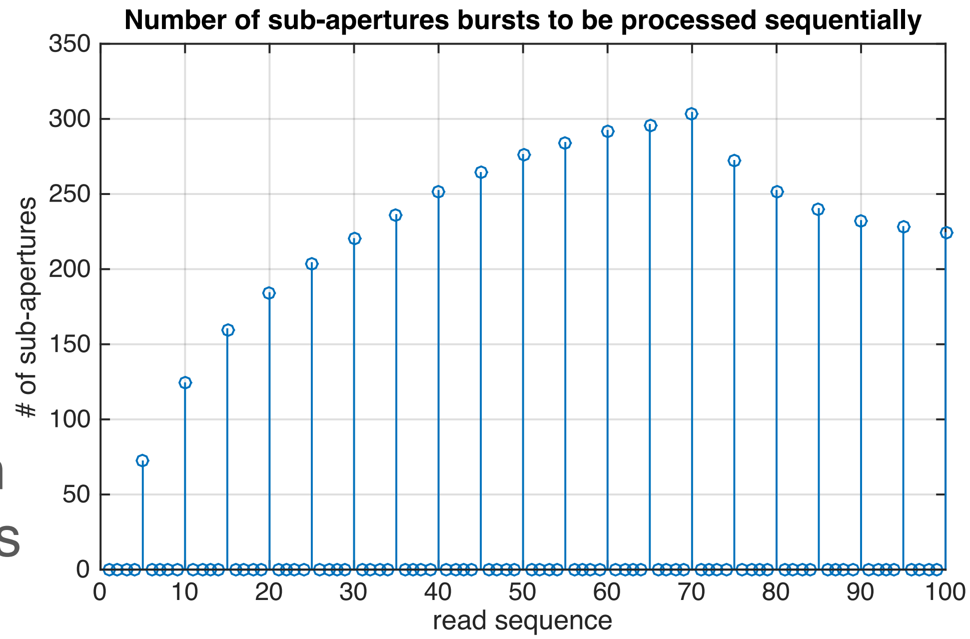
Benchmarkings

- 7250 Xeon Phi in flat mode quadrant memory
- 4 MPI tasks compatible with the future SNC-4 mode
- #1: POL slopes: ~900μs (with Intel KNL library)
    - If we only consider the reading of the matrix coefficients, the memory bandwidth is then equal to ~250 Gbytes/s. Considering the 450 Gbytes/s theoretical bandwidth, the result is coherent but we can hope to improve this result with a SNC-4 partition
- #2: POL + R
    - Global computation for the POL slopes
    - Stripe computation *nThreads:
    - Pixel calibration
    - Gradient computation
    - Reference subtraction
    - Projection for the 2 local slopes

- Timings
    - 4 threads: 1501μs
    - 16 threads: 1520μs
    - 32 threads: 1870μs
    - 40 threads: 2200μs.

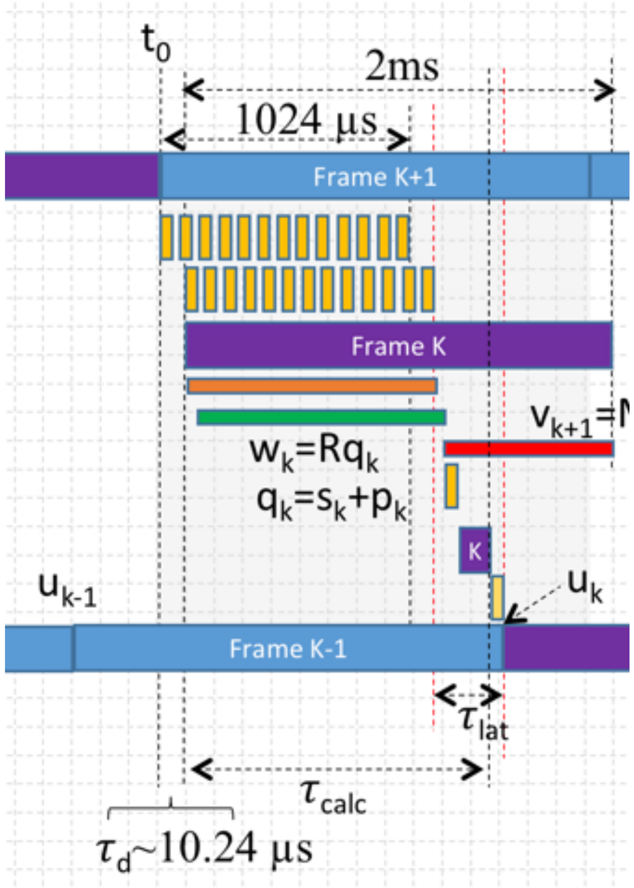
Benchmarkings

- #2: POL + R
  - Global computation for the POL slopes
  - Stripe computation *nThreads:
  - Pixel calibration
  - Gradient computation
  - Reference subtraction
  - Projection for the 2 local slopes
- Memory is allocated using a first-touch allocation policy, the nearest memory is used and the affinity between task and memory is then correct.
- Timings~1870µs,
  - POL ~ 800µs
  - Reconstruction ~1070µs.



RTC latency: summary

- Latency
  - Tomographic reconstruction latency ~50µs
  - Vector sending
  - TF node latency ~40µs
  - Vector sending
- 1 node for the POL+R
  - Total estimated latency ~340µs.



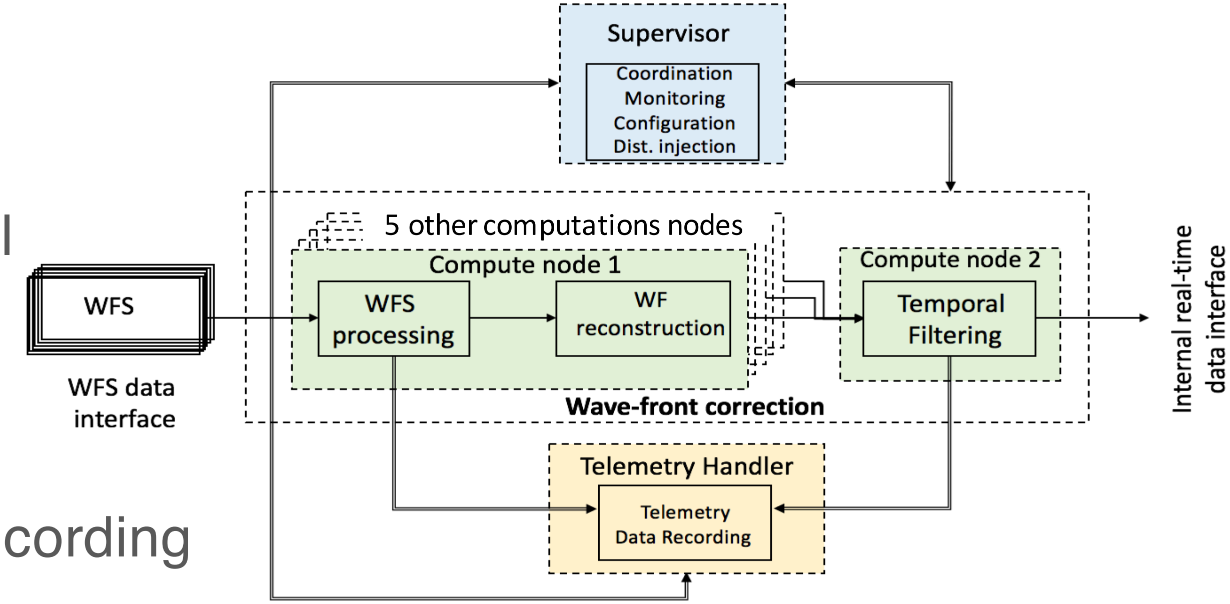
Improvements to *pipelinability*

- Alternative calculation procedures or redistribution of calculations differently amongst nodes.
  - swapping the POL with the Tomography step may be a fallback solution for the case $\delta \geq 1$ in case the latency required by the reconstruction proves too large on account of the variable bandwidth requirements
  - Off-load part of the POL computation to the TF node (or use 2 additional customised nodes: since M.vk is benchmarked to take ~850μs, the Time Filtering node could accommodate two such calculations with the loop time of 2ms with two more nodes for the remainder 4 M.vk calculations) POL is partially done in the TF node
    - Latency ~ 690 µs.
  - Use convolutive model to compute M.vk since the interaction matrix M can be well approximated by stencil operations for Shack-Hartmann WFS
  - Replace the POL by pre-computed R*M which is a 6k x 6x matrix instead of M which is a 6k x 60k matrix

$$w_k = R(s_k + Mp_k) \rightarrow w_k = Rs_k + RMp_k$$

Summary

- 1 node/WFS can handle Pixel Processing + Reconstruction
- 1 node for Time Filtering
- 1 node for supervision
- 1 node for Telemetry Data Recording
- Total=9 nodes
  - Latency: 340µs
  - Jitter: 100µs measured, 40µs (goal) needs further assessment
- SW architecture
  - parallel tasks exchanging messages and data across a 10 Gbits Ethernet network



Questions?

ACKNOWLEDGEMENTS

The research leading to these results received the support of the A*MIDEX project (no. ANR-11-IDEX-0001- 02) funded by the ”Investissements d'Avenir” French Government program, managed by the French National Research Agency (ANR)