POST-SIEVING ON GPUs

A. Ma [1],
J. W. Bo [2],
T. Ka [1],
A. K. La [1]

[1] LACAL, EPFL, La, Sw
[2] NXP Se, La, Be

NUMBER FIELD SIEVE (NFS)

- A [illegible]
- RSA 768-bit [illegible] NFS in 2010
- [illegible]

$$x, y : x^2 \equiv y^2 \bmod n \text{ and } x \not\equiv \pm y \bmod n$$

- Two steps
  RELATION COLLECTION: [illegible] ≈90%T
  LINEAR ALGEBRA STEP: [illegible] ≈10%T

NFS RELATIONS

- T[illegible] B[illegible] $_r$, B_a
- I[illegible] $f_r(X)$, $f_a(X)$ f[illegible] a[illegible] (=5,6)

- R[illegible] (a,b) t[illegible] a,b p[illegible] g(b>0) s[illegible]

1. b $_r$(a/b) s[illegible] B $_r$-s[illegible] ≤ 3 p[illegible] B $_r$ a[illegible] ≤ B_L
2. b^{d}f$_a$(a/b) s[illegible] B $_a$-s[illegible] ≤ 4 p[illegible] B $_a$ a[illegible] ≤ B_L

COLLECT RELATIONS

SIEVING: find (a, b) such that $f_r(a, b)$ and $(b^d f_a(a, b))$ is smooth over B_r - basis (B_a - basis) prime "basis" for $\leq B_L^3$ (B_L^4)

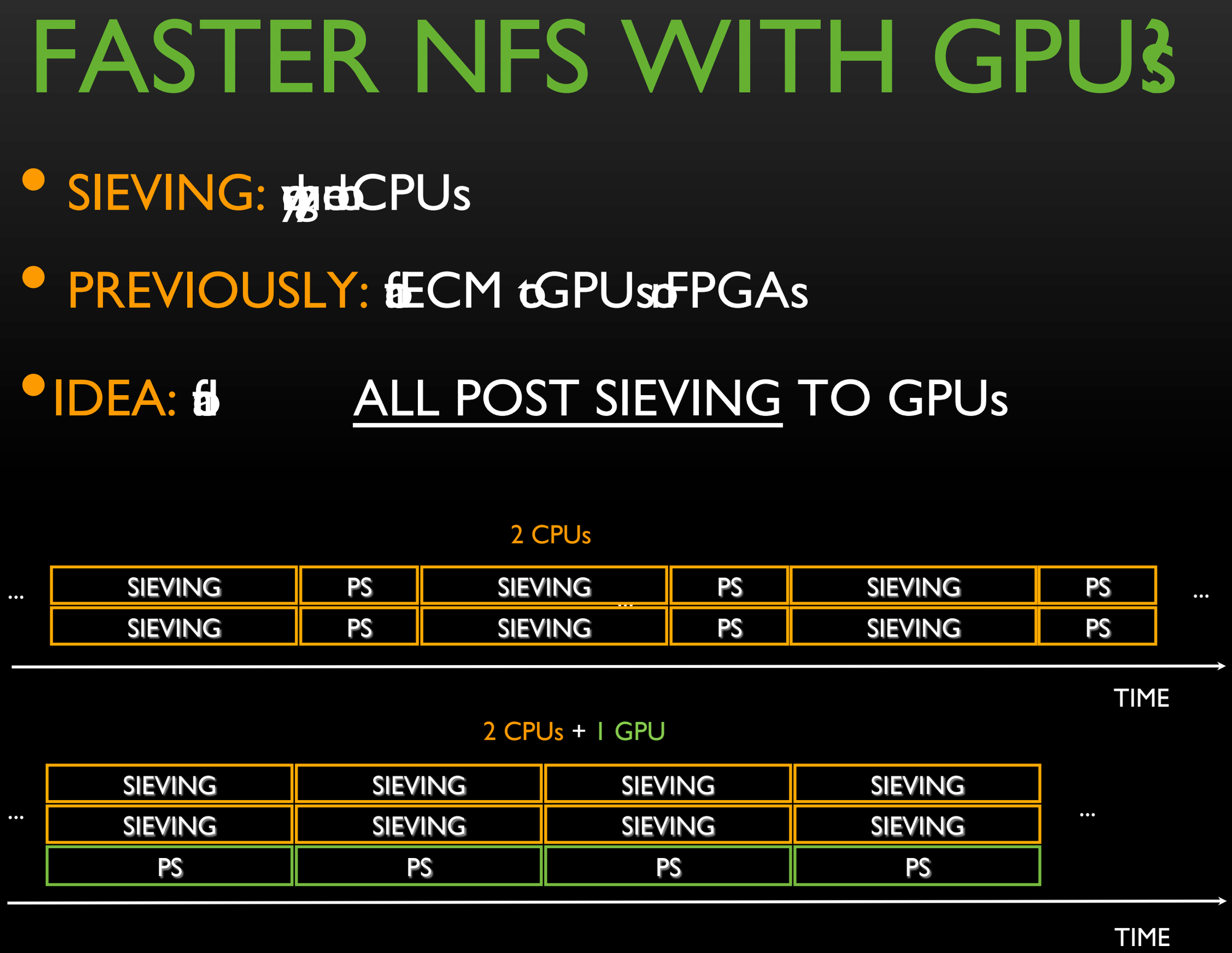
POST SIEVING (NORMALLY 12-17% OF THE TOTAL TIME):

1. Compute $f_r(a, b)$ and $b^d f_a(a, b)$
2. Remove duplicate relations
3. Factor $f_r(a, b)$ and $b^d f_a(a, b)$ (COFACTORING)

EMBARRASSINGLY PARALLEL!

FASTER NFS WITH GPUs?

- SIEVING: [illegible] CPUs
- PREVIOUSLY: [illegible] ECM to GPUs or FPGAs
- IDEA: [illegible] ALL POST SIEVING TO GPUs

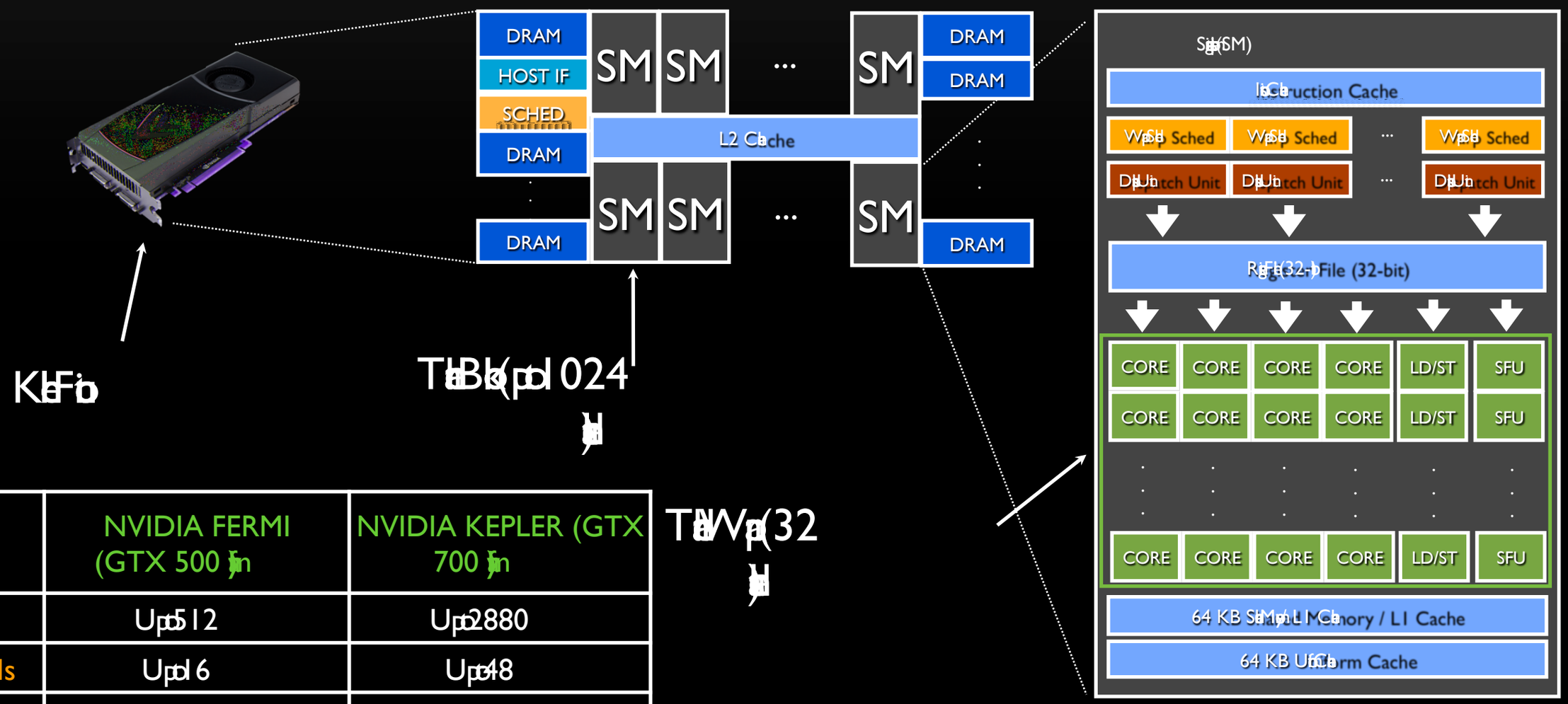


GPUs NOT ONLY GAMING…

M[illegible]32-[illegible]GPGPU, [illegible]

O[illegible]

W[illegible].



	NVIDIA FERMI (GTX 500 [illegible]	NVIDIA KEPLER (GTX 700 [illegible]
Co[illegible]	Up[illegible]512	Up[illegible]2880
SMs	Up[illegible]16	Up[illegible]48
Fr[illegible]	Up[illegible]544 MHz	Up[illegible]980 MHz
DRAM	Up[illegible]3GB (192 GB/[illegible]	Up[illegible]6 GB (336 GB/[illegible]

COFACTORING ON GPUsOUTLOOK

- Ip
- Op
- Tw CUDA Kl

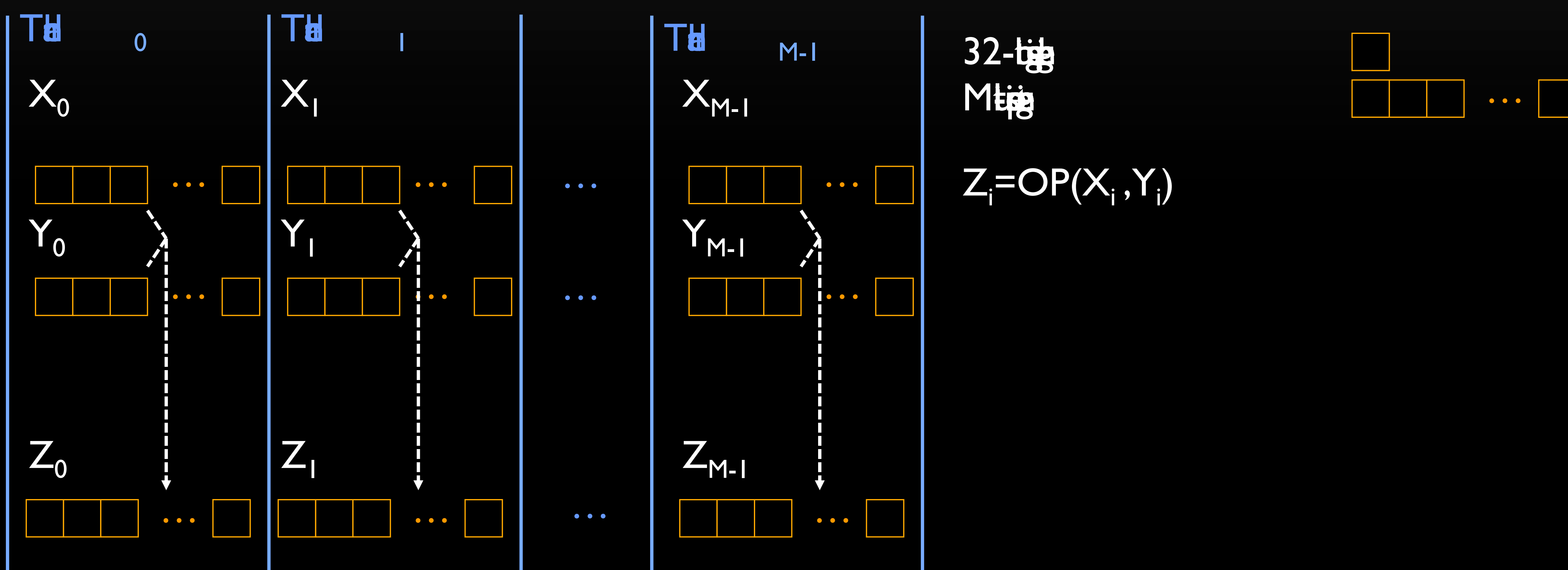
1. R
2. A

DESIGN STRATEGY

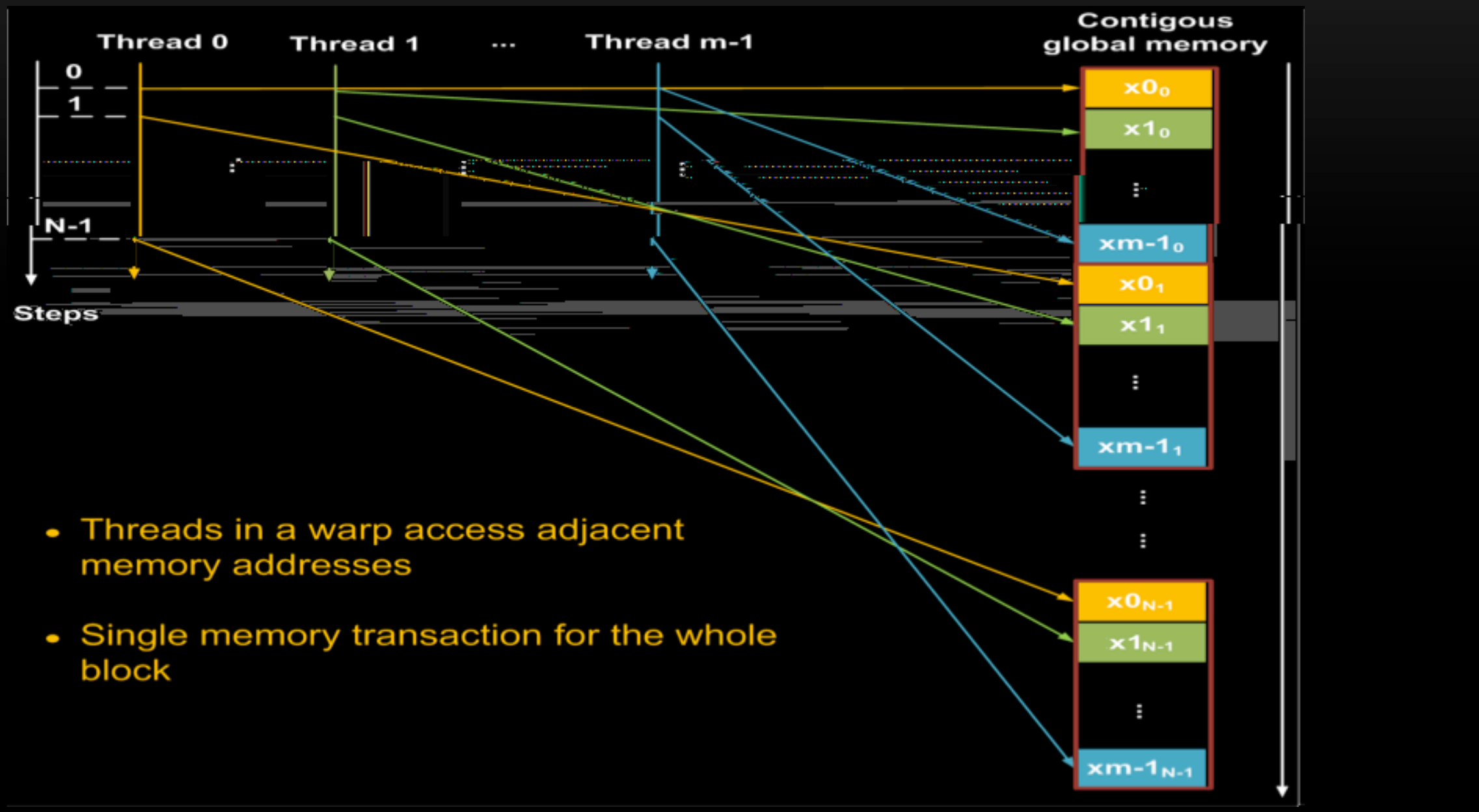
ARITHMETIC DESIGN

S… R… 2 ^{32} M…n

PTX … (MAD …)



GLOBAL MEM ACCESS: COALESCING



KERNEL ANATOMY

PREAMBLE

1. R
2. R

F

COFACTOR FACTORIZATION: REPEAT K TIMES

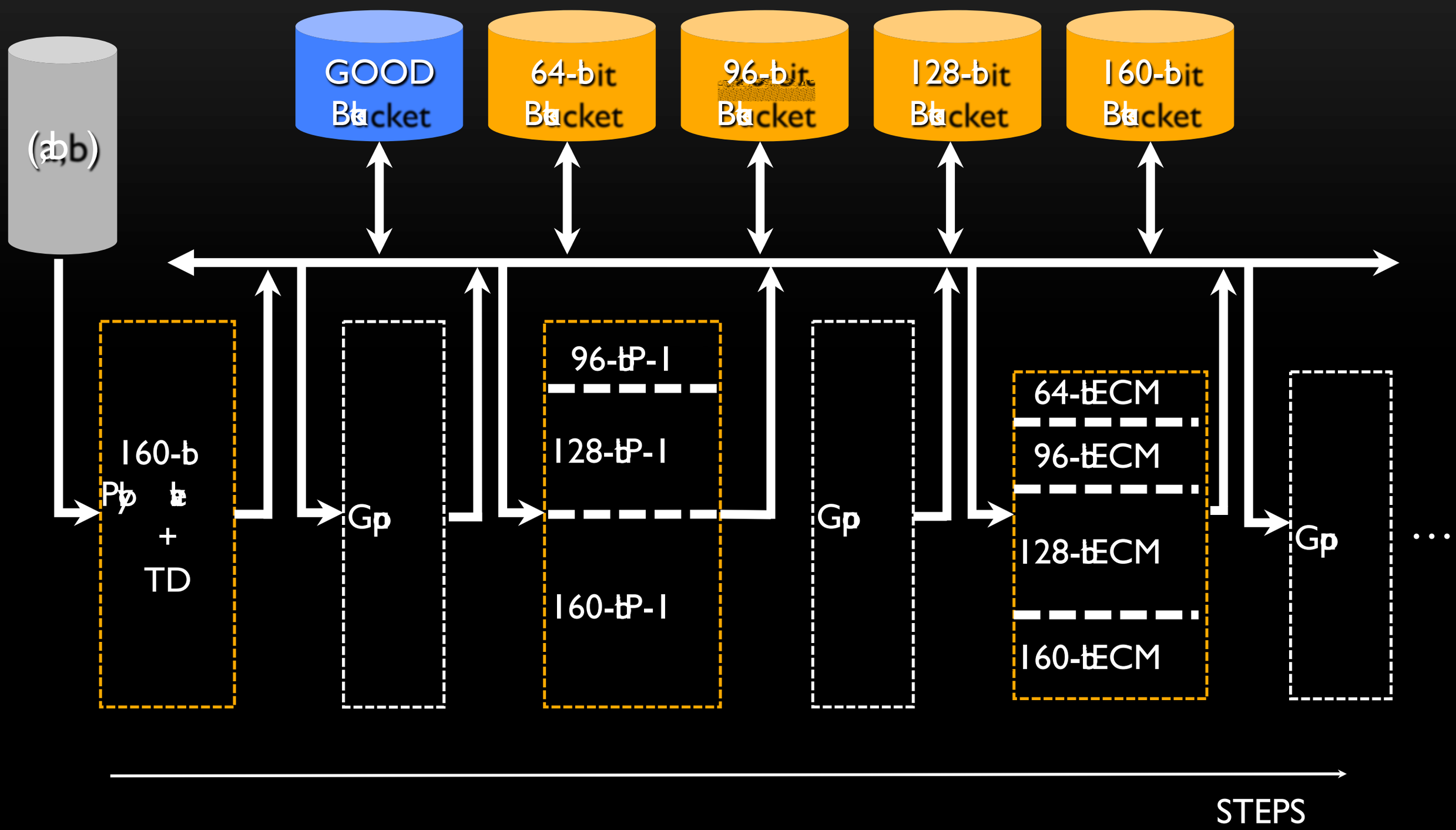
1. G

2. F E C M (

3. I

4. D B $\leq B_L$

KERNEL WORKFLOW



ABOUT THE ALGORITHMS…

- Bi
  jd He
- TD
  pCMEM, (H/M)
- P
  SgRaMe
- P P-1 (M
  , pBSGS g2
- ECM (M
  T , pBSGS g2

INTEGRATION WITH RSA-768 SOFTWARE

Fi[illegible] GPU [illegible]

- P[illegible]
- We[illegible] RSA-768 [illegible]

W[illegible]

- We[illegible]
- F[illegible] 95% [illegible] 99% [illegible]
- T[illegible]

RSA 768: CHOICE OF PARAMETERS

- Very large B_1 and B_2: $B_1 \approx 2^{10}$, $B_2 \approx 2^{14}$
- 8-20 ECM curves per B_1: $B_1 = [2^8, 2^{10}]$, $B_2 = [2^{12}, 2^{15}]$

CPU vs GPU

CPU: INTEL I7-3770K 4 @ 3.5 GHz 16GB RAM

Lag in ps	Inputs	Time in CPU	Size in CPU	PS-6 in CPU	Rho found
≤ 3	$\approx 5 \times 10^5$	29.6s	25.6s	4.0s	125
≤ 4	$\approx 10^6$	32.0s	25.9s	6.1s	137

GPU: NVIDIA GTX 580 512 CORES 1544 MHz 1.5 GB RAM

Lag in ps	Inputs	Density	CPU/GPU Ratio	Time	Rho found
≤ 3	$\approx 5 \times 10^5$	95%	9.8	2.6s	132
		99%	6.9	3.7s	136
≤ 4	$\approx 10^6$	95%	4.0	6.5s	159
		99%	2.7	9.6s	165

1 CPU vs 1 CPU + 1 GPU

Lg jn	# Ip jp	Sg	Tn	# Rd fd	Rs
≤ 3	≈ 5 0 ^{7}	NGPU	2961s	12523	4.23
		WtGPU	2564s	13761	5.37
≤ 4	≈ 5 0 ^{7}	NGPU	1602s	6855	4.28
		WtGPU	1300s	8302	6.39

Lgn ≤ 3 24% GAIN
Lgn ≤ 45% GAIN

CONCLUSIONS AND FUTURE WORK

GPU
The
We

Op
Ge

NVIDIA
RSA 1024-b

Kp

GPUs(AMD?)