High Performance Linkage Disequilibrium: FPGAs Hold the Key

Nikolaos Alachiotis and Gabriel Weisz

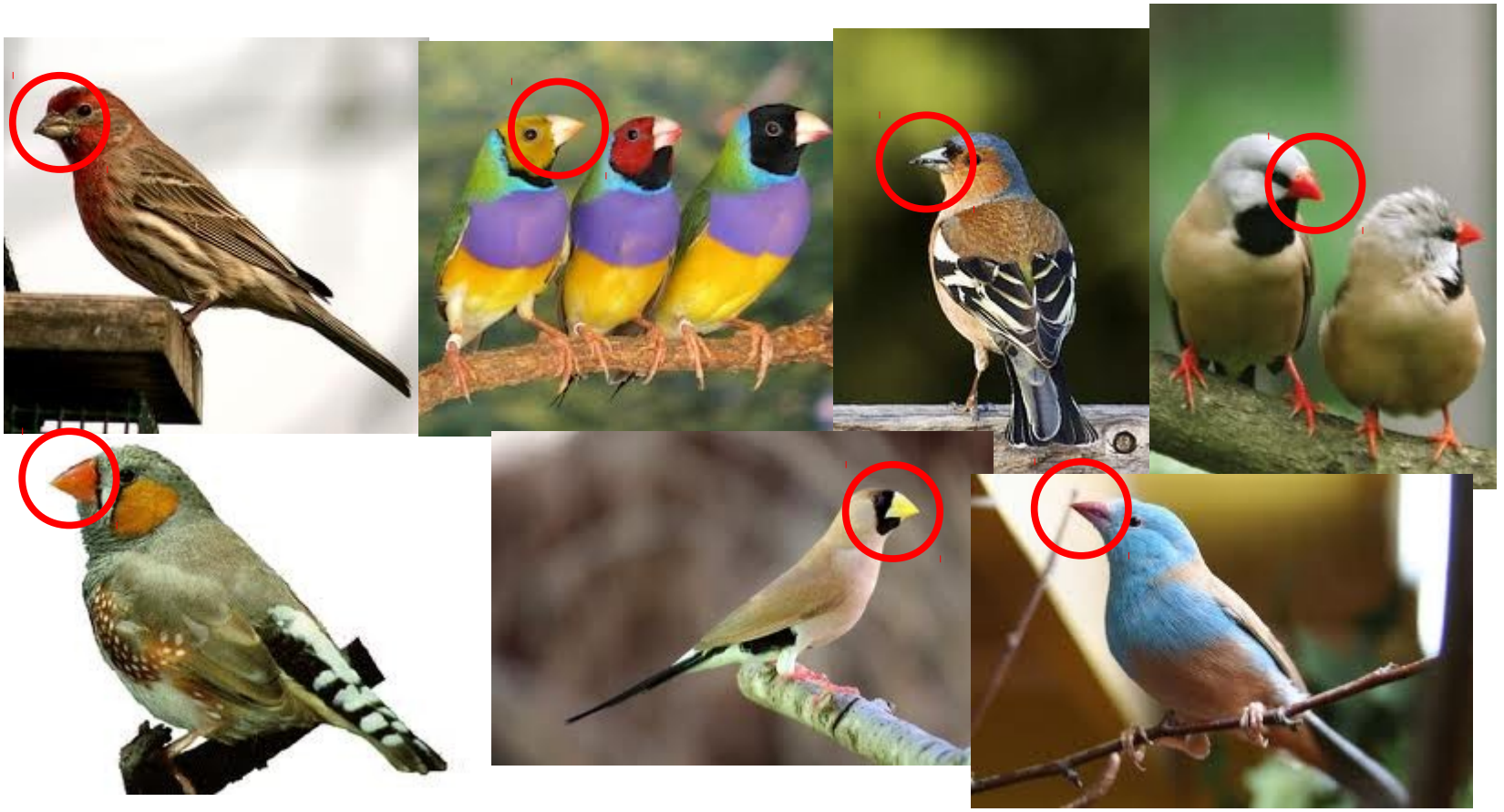
Department of Electrical & Computer Engineering
Carnegie Mellon University

Outline

- Motivation
- “Generic” algorithm for Linkage Disequilibrium (LD)
- Reconfigurable accelerator architecture and hardware generation
- Design space exploration and performance comparison
- Conclusion

Population genetics investigate...

… the adaptation of species in an environment by studying the genetic composition of a population (same species).



A finch collection

The shape of the beak is controlled by genes.

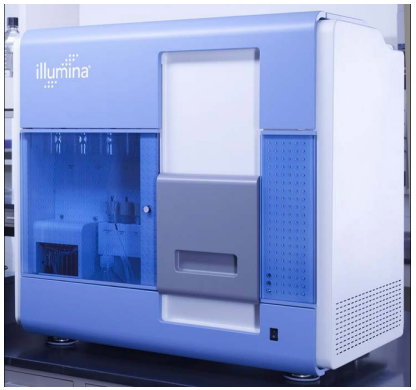
What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.

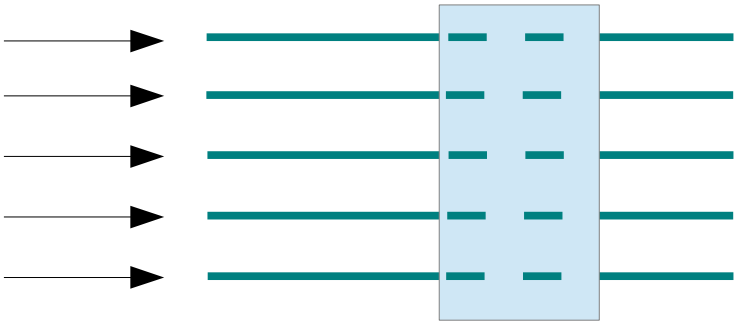
What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.

DNA sequencing

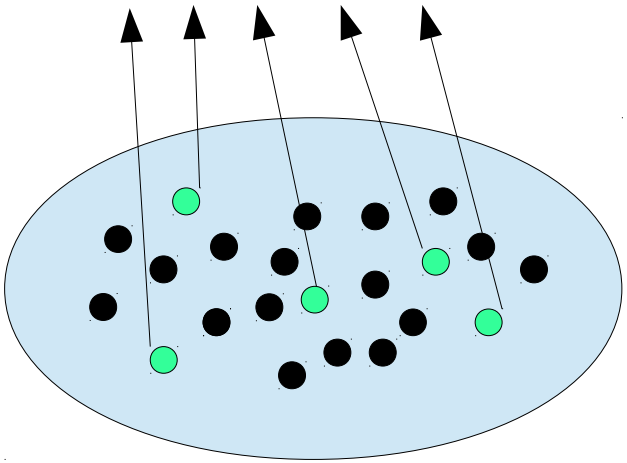


Multiple Sequence Alignment (MSA)



Genomic region

A C C C G C A C T
A C C C G C A G T
A C T C A C C C T
A C C C A C C C T
C C C C A C C C T

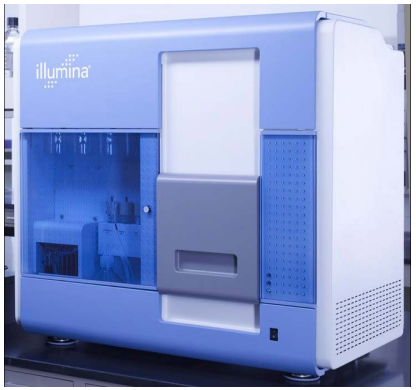


Population

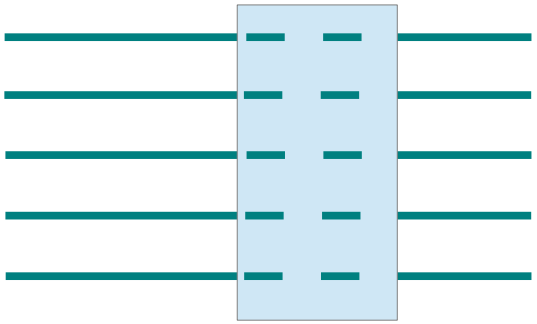
What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.

DNA sequencing



Multiple Sequence Alignment (MSA)

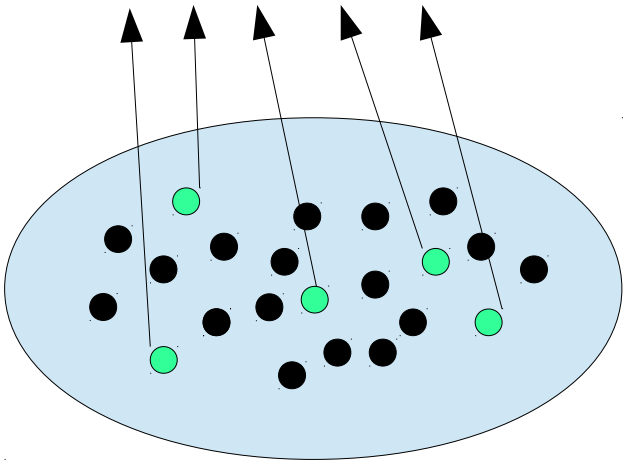


Outgroup

A C C C G C A C T

A C C C G C A C T
A C C C G C A G T
A C T C A C C C T
A C C C A C C C T
C C C C A C C C T

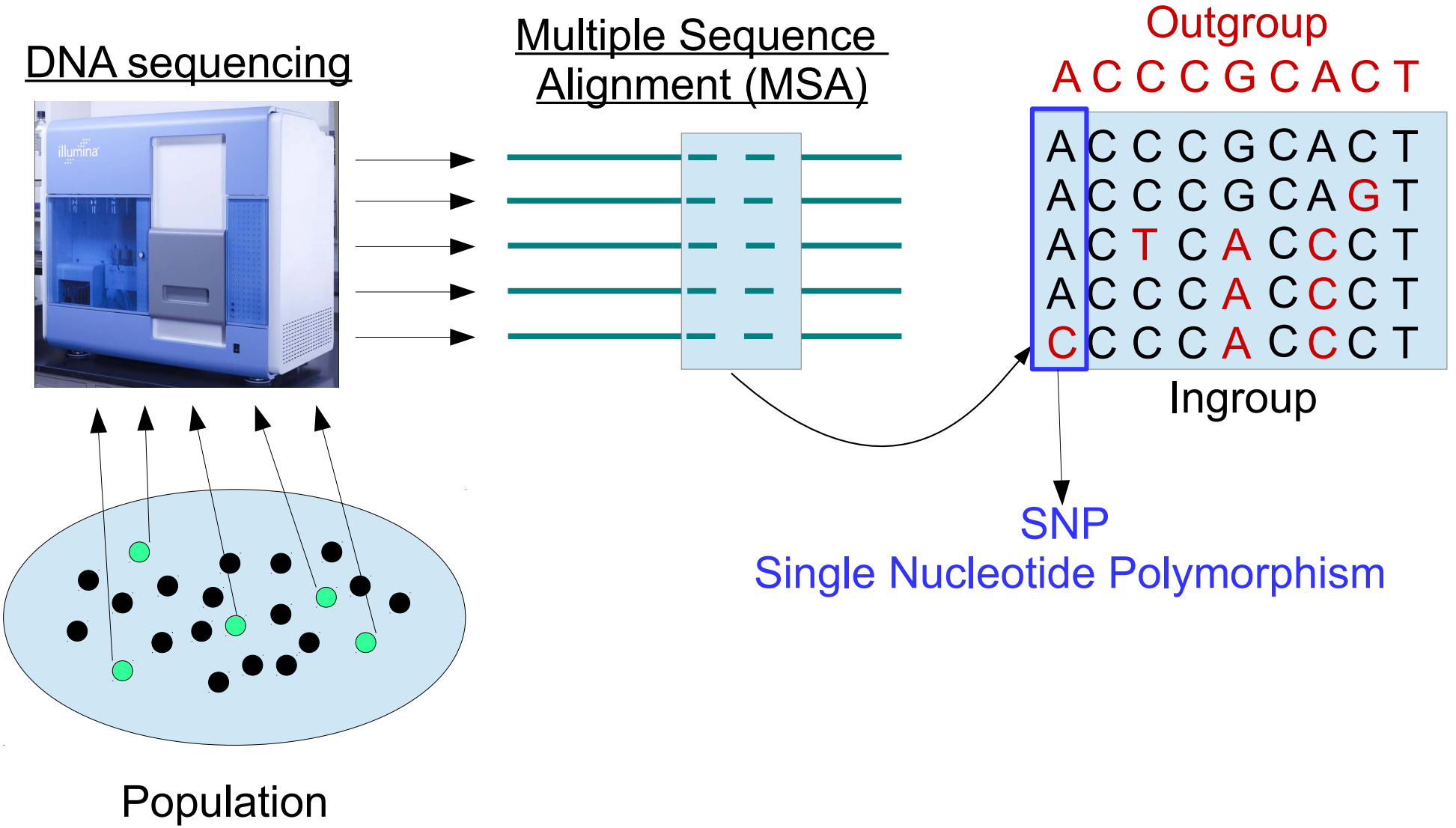
Ingroup



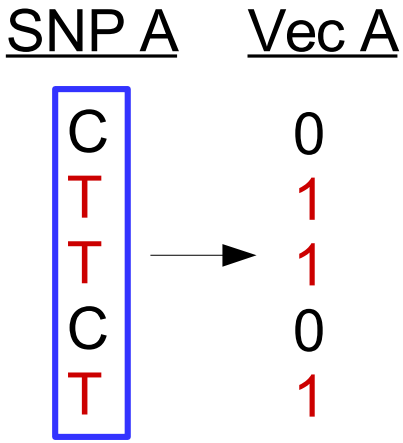
Population

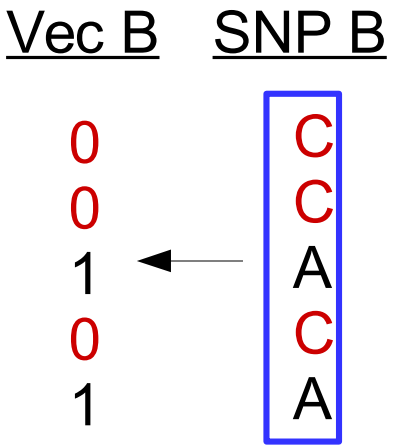
What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.



Pairwise LD computation





Alleles: C and T

Ancestral: C (outgroup) → 0

Derived: T → 1

Alleles: C and A

0 ← Ancestral: A

1 ← Derived: C

Pairwise LD computation

Derived allele frequencies

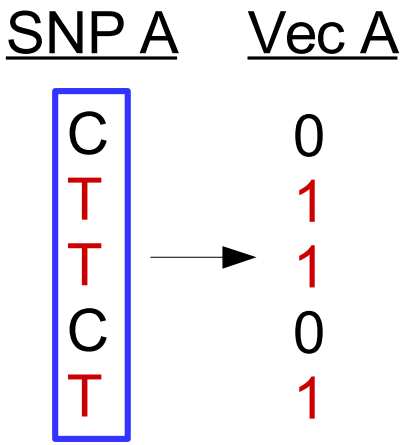
$$p_1 = popcnt(A) \div N$$

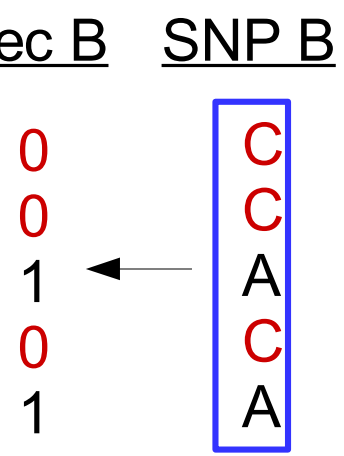
$$q_1 = popcnt(B) \div N$$

Haplotype frequency

$$x_{11} = popcnt(A \cap B) \div N$$

N: number of genomes





Alleles: C and T

Ancestral: C (outgroup) → 0

Derived: T → 1

Alleles: C and A

0 ← Ancestral: A

1 ← Derived: C

Pairwise LD computation

Derived allele frequencies

$$p_1 = popcnt(A) \div N$$

$$q_1 = popcnt(B) \div N$$

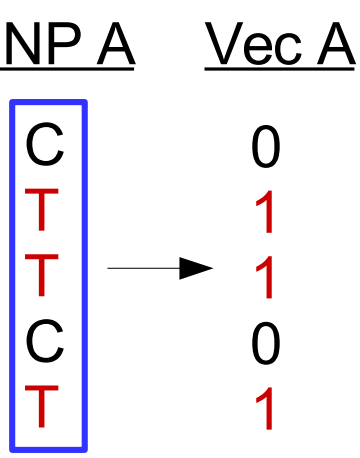
Haplotype frequency

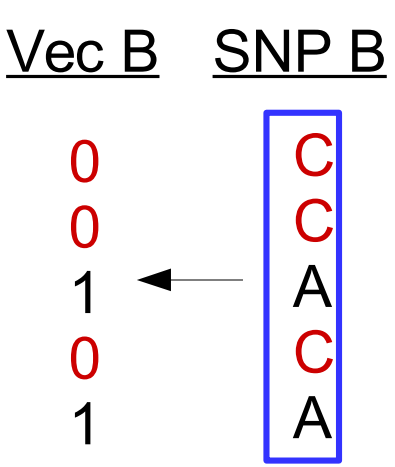
$$x_{11} = popcnt(A \cap B) \div N$$

N: number of genomes

Squared Pearson correlation coefficient

$$r^2_{AB} = \frac{(x_{11} - p_1 q_1)^2}{p_1 q_1 (1 - p_1)(1 - q_1)}$$





Alleles: C and T

Ancestral: C (outgroup) → 0

Derived: T → 1

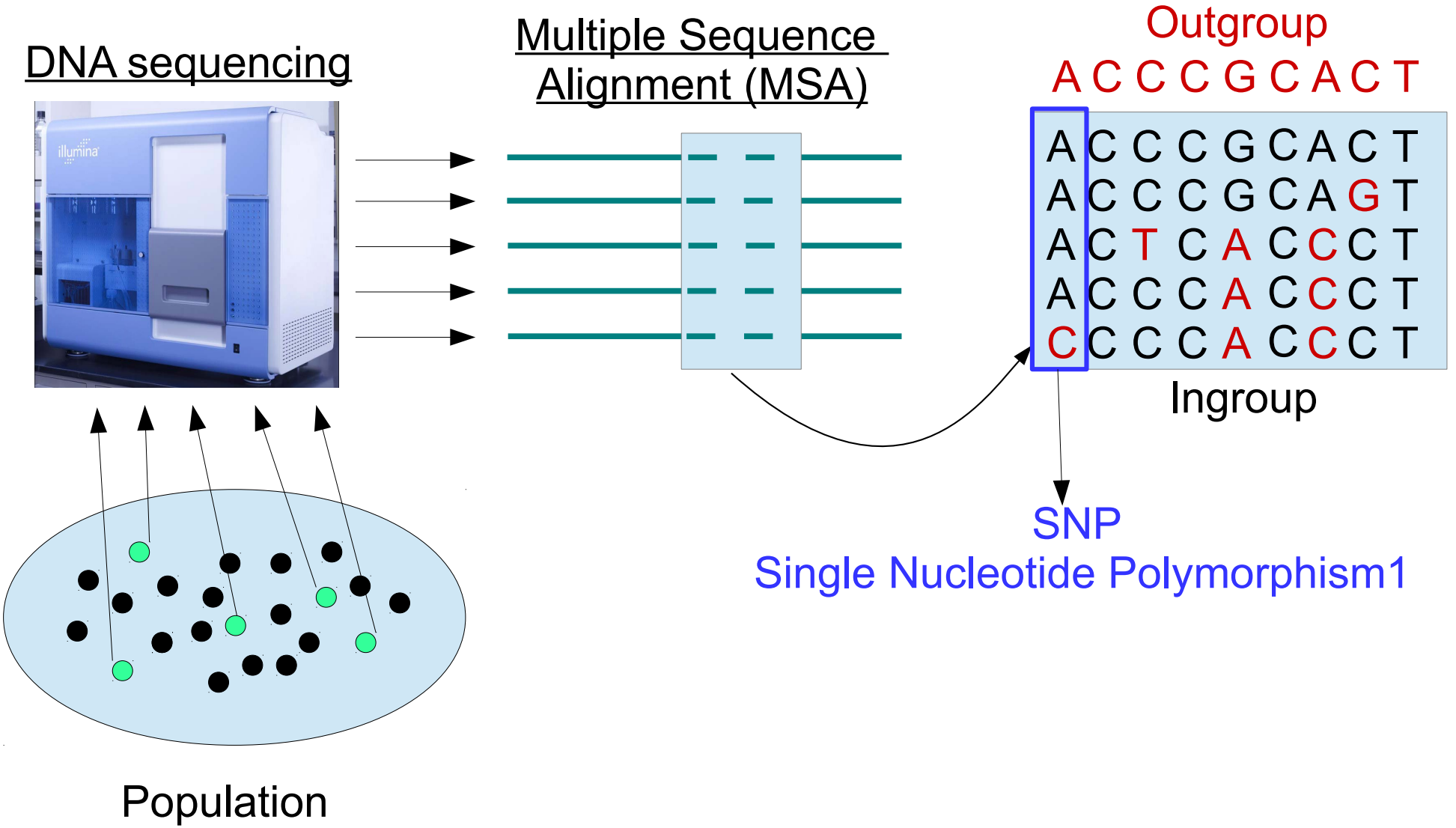
Alleles: C and A

0 ← Ancestral: A

1 ← Derived: C

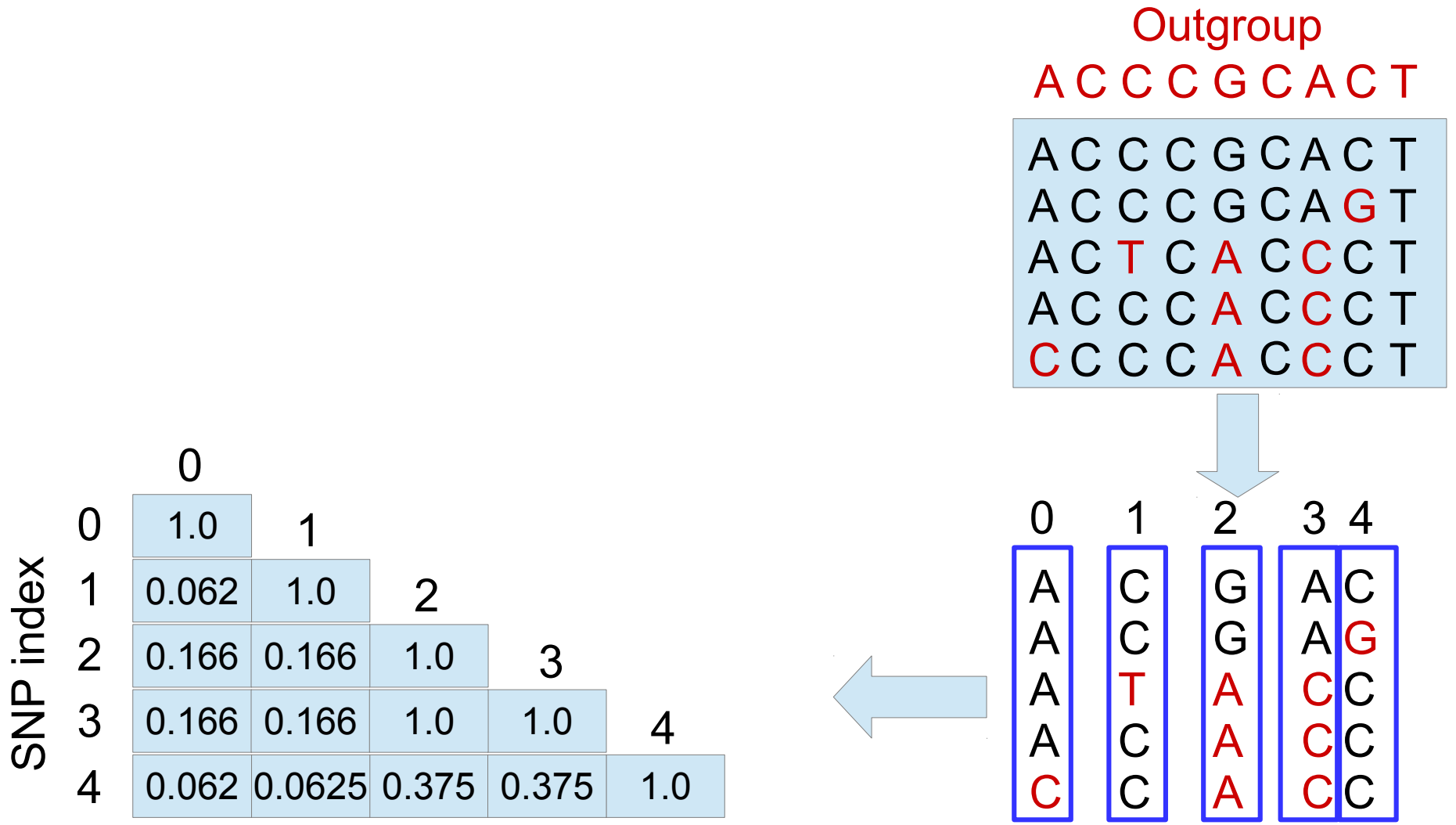
What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.



What is Linkage Disequilibrium (LD)

- LD is the non-random association between alleles (different forms of a gene) at different locations in a genome.



LD scores based on Pearson's coefficient of correlation

Applications of LD

- Identification of traces of positive selection

GENETICS

Linkage Disequilibrium as a Signature of Selective Sweeps

Yuseob Kim, Rasmus Nielsen

GENETICS *July 27, 2004 vol. 167 no. 3 1513-1524; DOI: 10.1534/genetics.103.025387*

- Disease gene identification

nature International weekly journal of science

Letters to Nature

Nature **411**, 199-204 (10 May 2001) | doi:10.1038/35075590; Received 11 December 2000; Accepted 13 March 2001

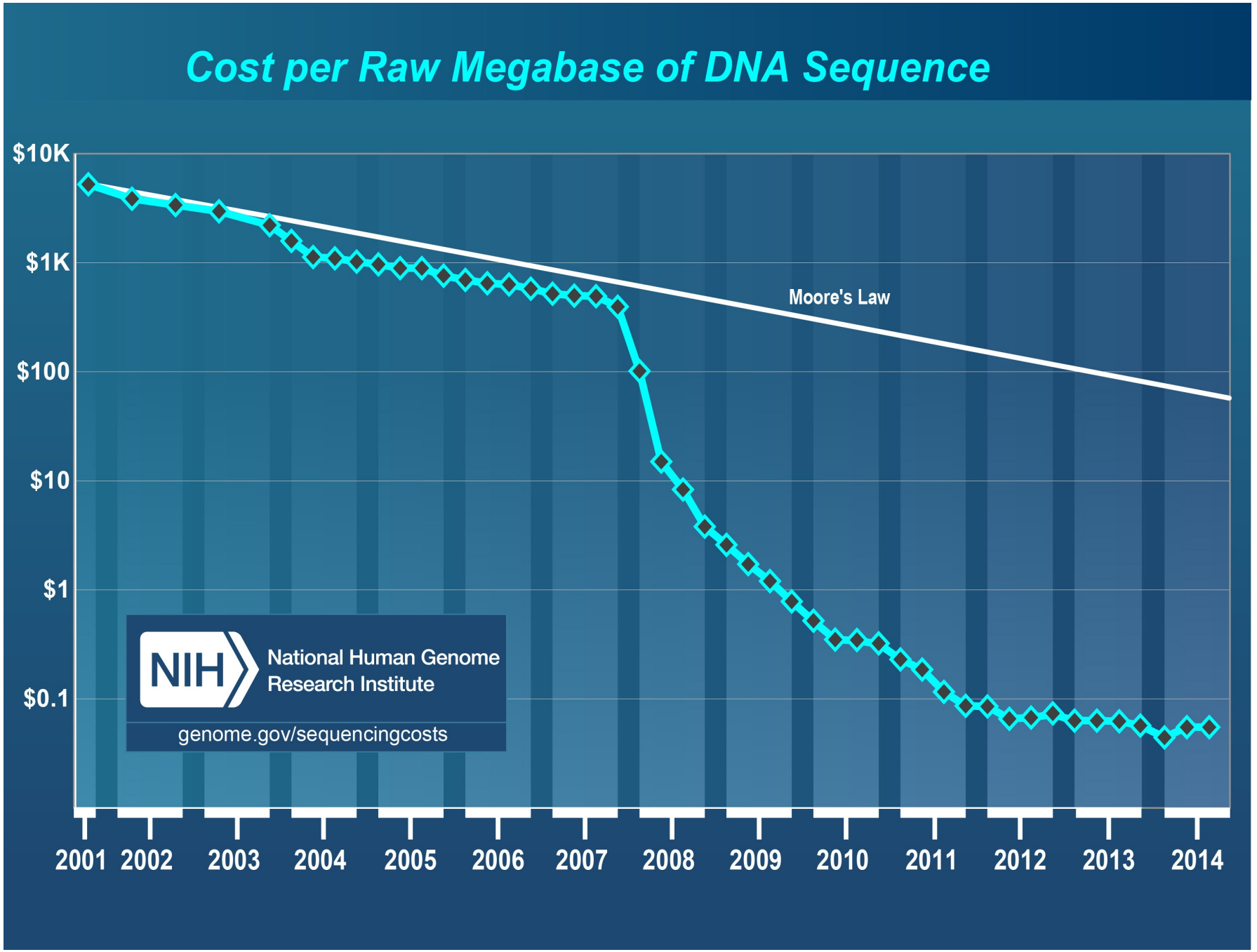
Linkage disequilibrium in the human genome

David E. Reich[1], Michele Cargill[1,2], Stacey Bolk[1], James Ireland[1], Pardis C. Sabeti[3], Daniel J. Richter[1], Thomas Lavery[1], Rose Kouyoumjian[1], Shelli F. Farhadian[1], Ryk Ward[3] & Eric S. Lander[1,4]

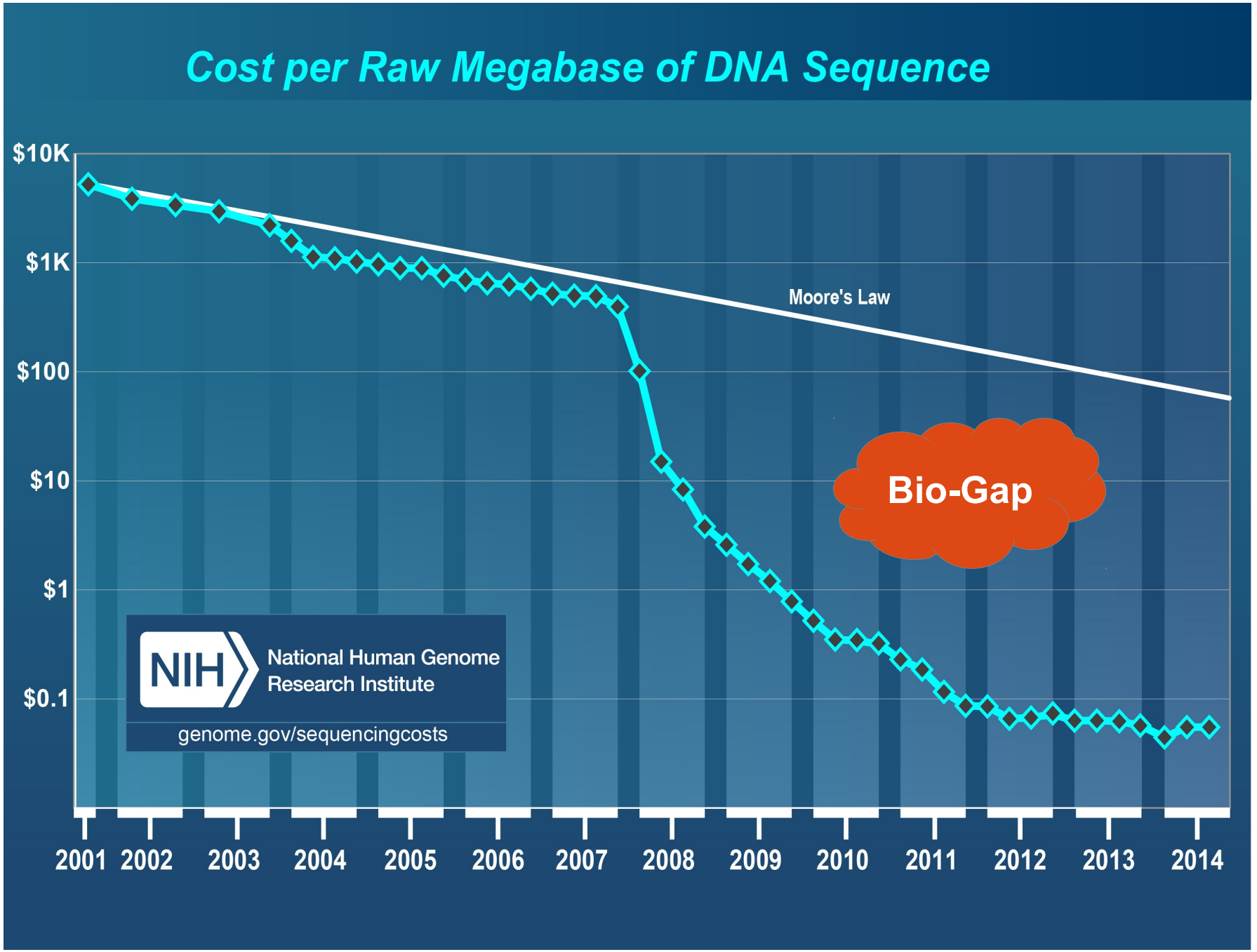
LD

So what are the computational issues with LD?

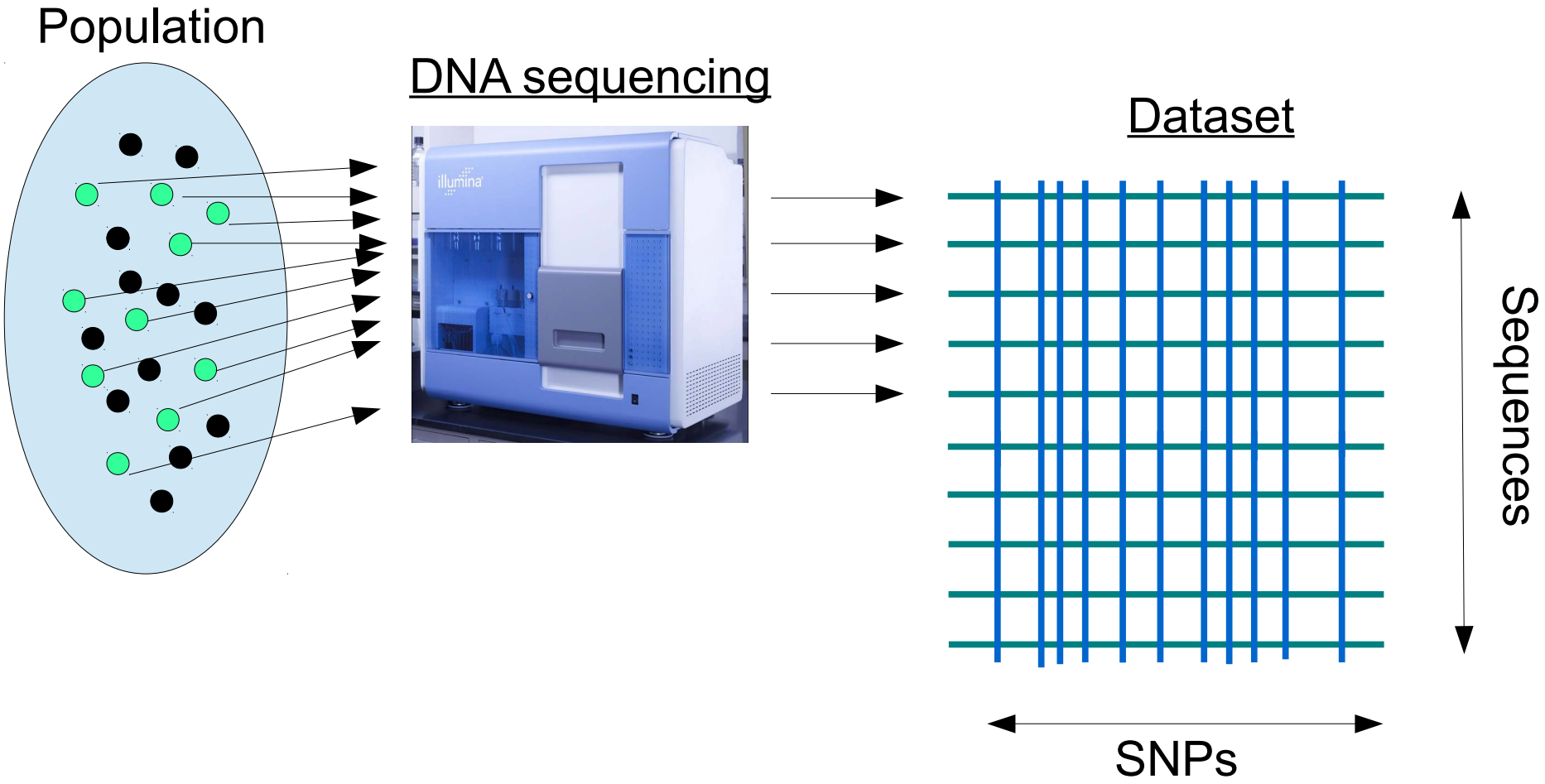
Advances in DNA Sequencing



Advances in DNA Sequencing

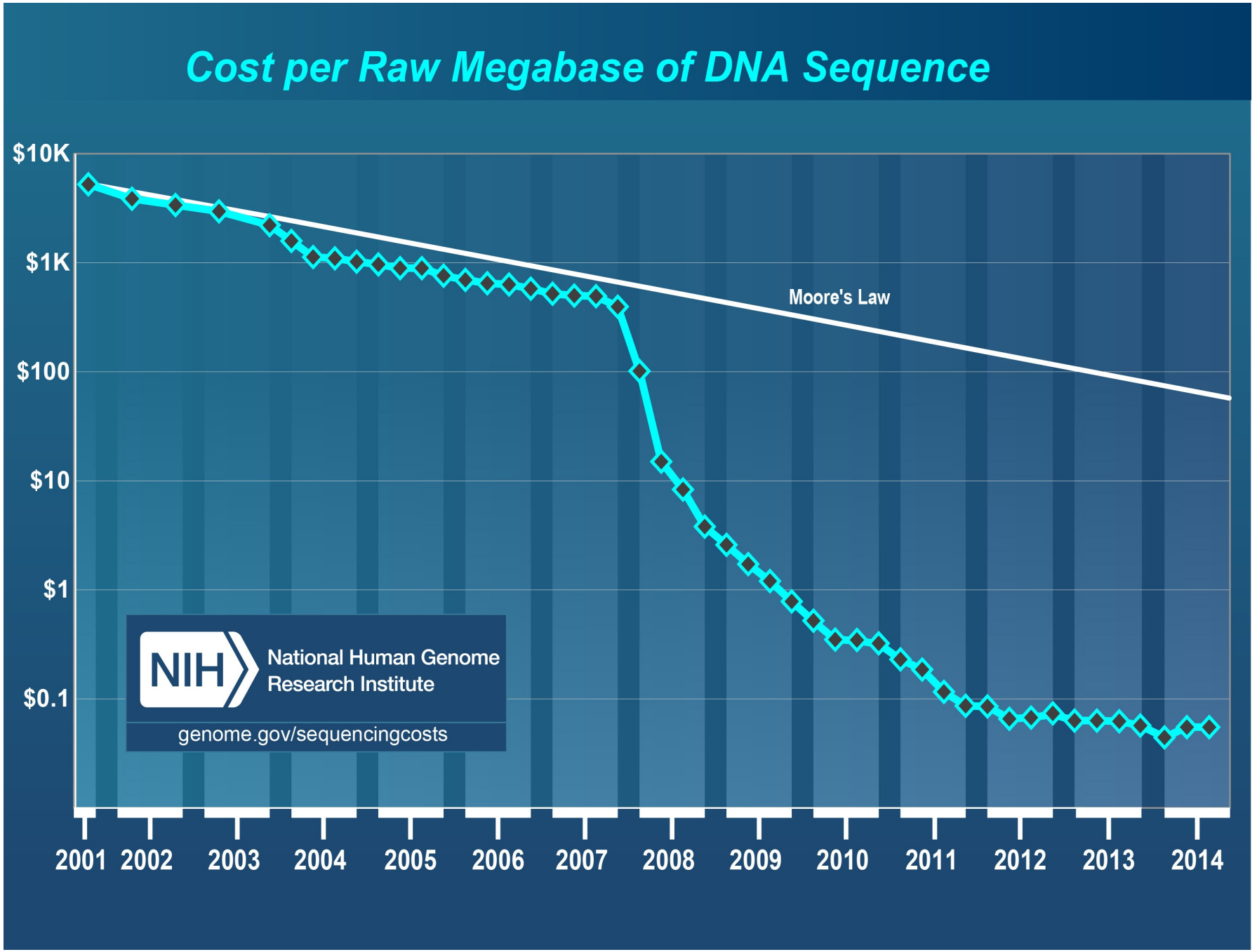


SNPs increase (almost) proportionally with the genomes



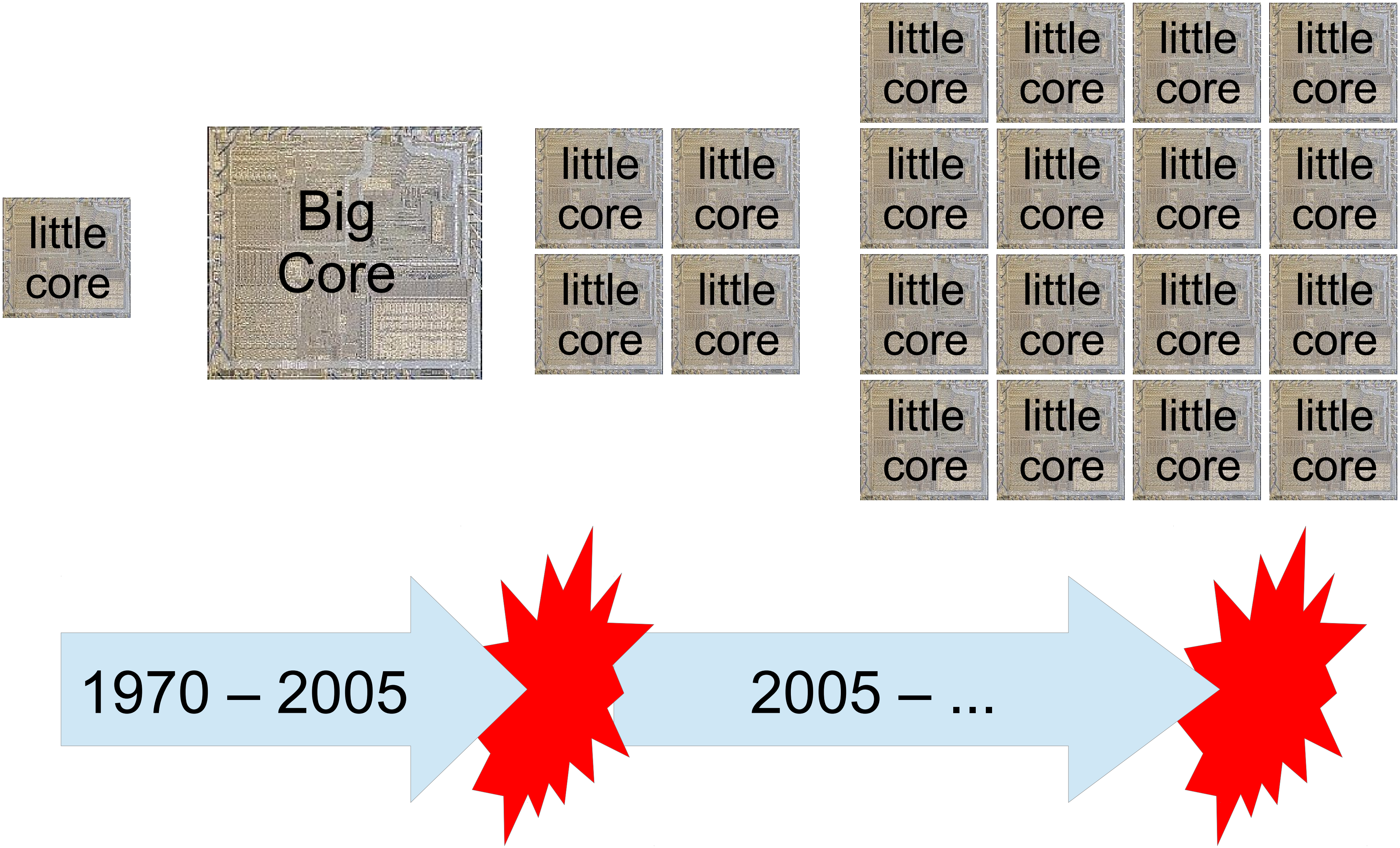
- Computational demands increase linearly with the number of genomes
- … but quadratically with the number of SNPs

In need for high performance



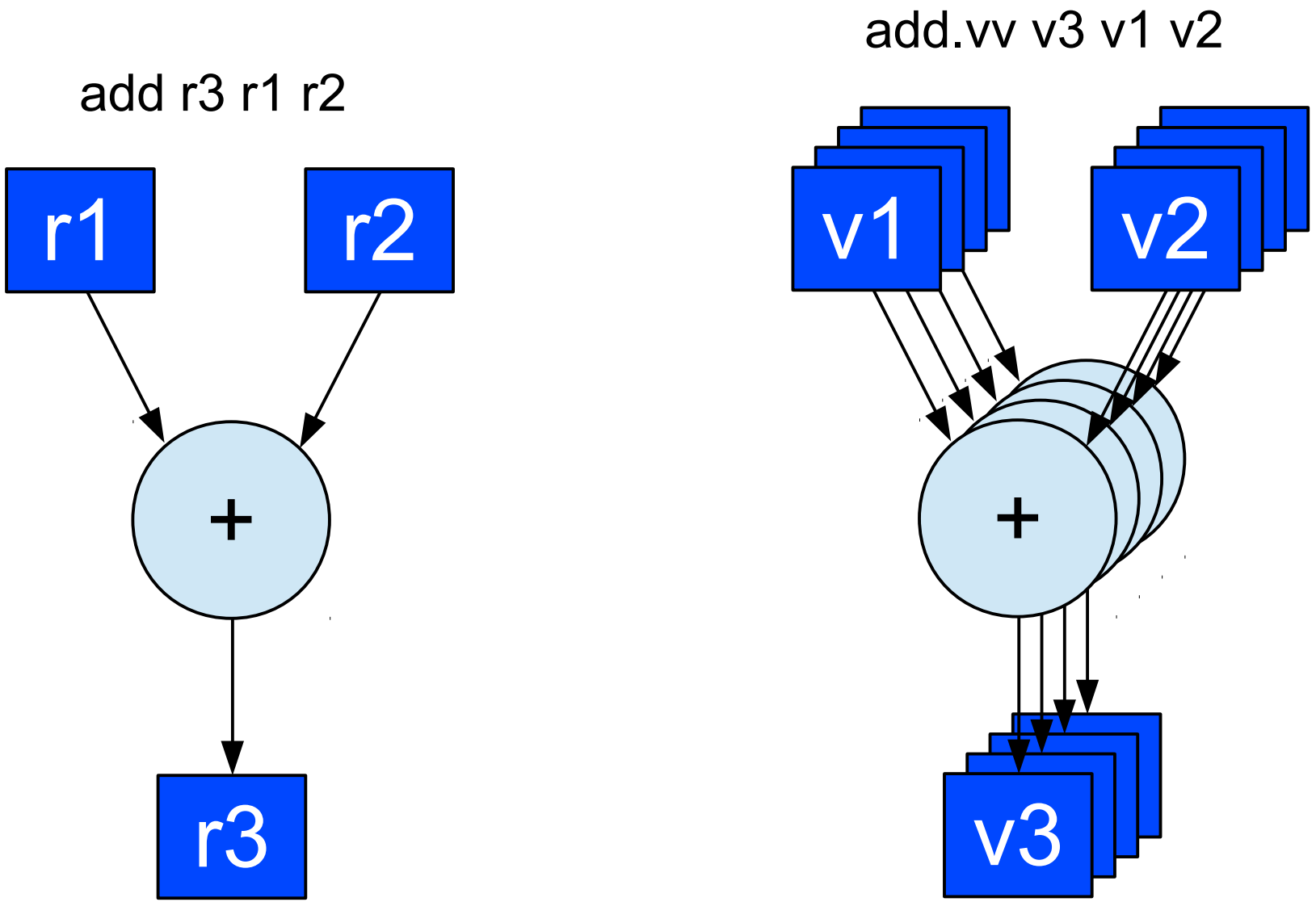
Moore's law

From single-core to multi-core processors



Moore's law

From scalar to vector processing



Performance bottleneck

Derived allele frequencies

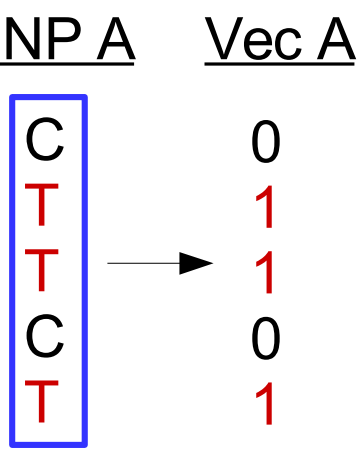
$$p_1 = popcnt(A) \div N$$
$$q_1 = popcnt(B) \div N$$

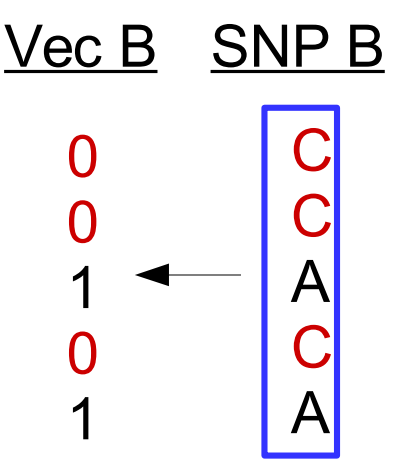
Haplotype frequency

$$x_{11} = popcnt(A \cap B) \div N$$

Squared Pearson correlation coefficient

$$r^2_{AB} = \frac{(x_{11} - p_1 q_1)^2}{p_1 q_1 (1 - p_1)(1 - q_1)}$$





Performance bottleneck

Derived allele frequencies

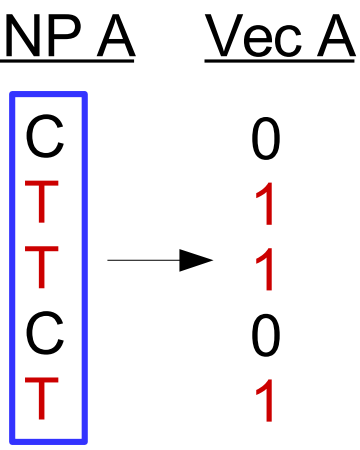
$$p_1 = popcnt(A) \div N$$
$$q_1 = popcnt(B) \div N$$

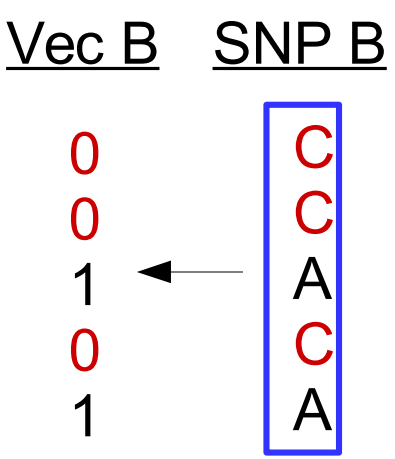
Haplotype frequency

$$x_{11} = popcnt(A \cap B) \div N$$

Squared Pearson correlation coefficient

$$r^2_{AB} = \frac{(x_{11} - p_1 q_1)^2}{p_1 q_1 (1 - p_1)(1 - q_1)}$$

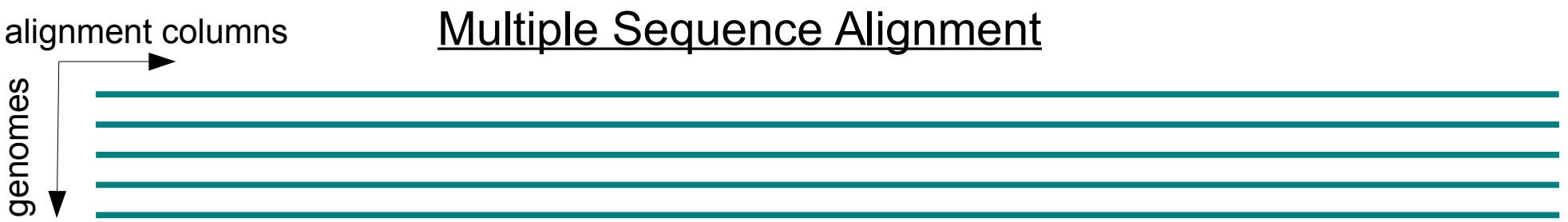




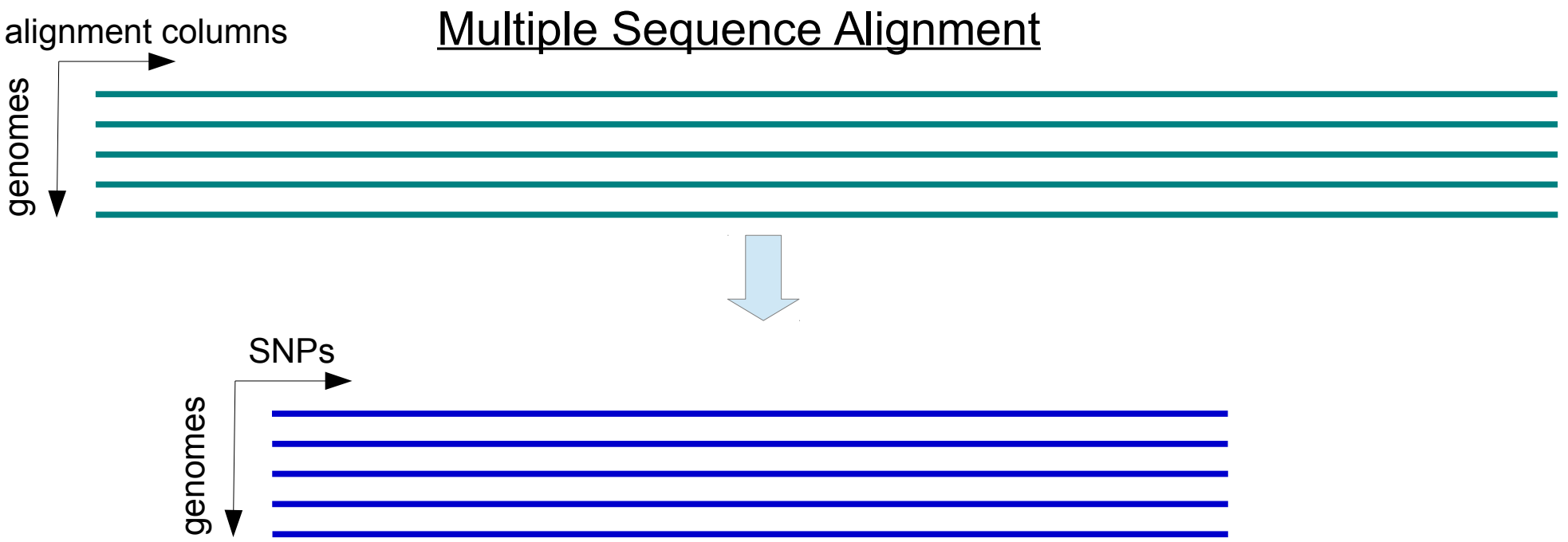
Hardware support (intrinsic instruction) for population count in processors but **only on regular registers.**

“Generic” LD approach

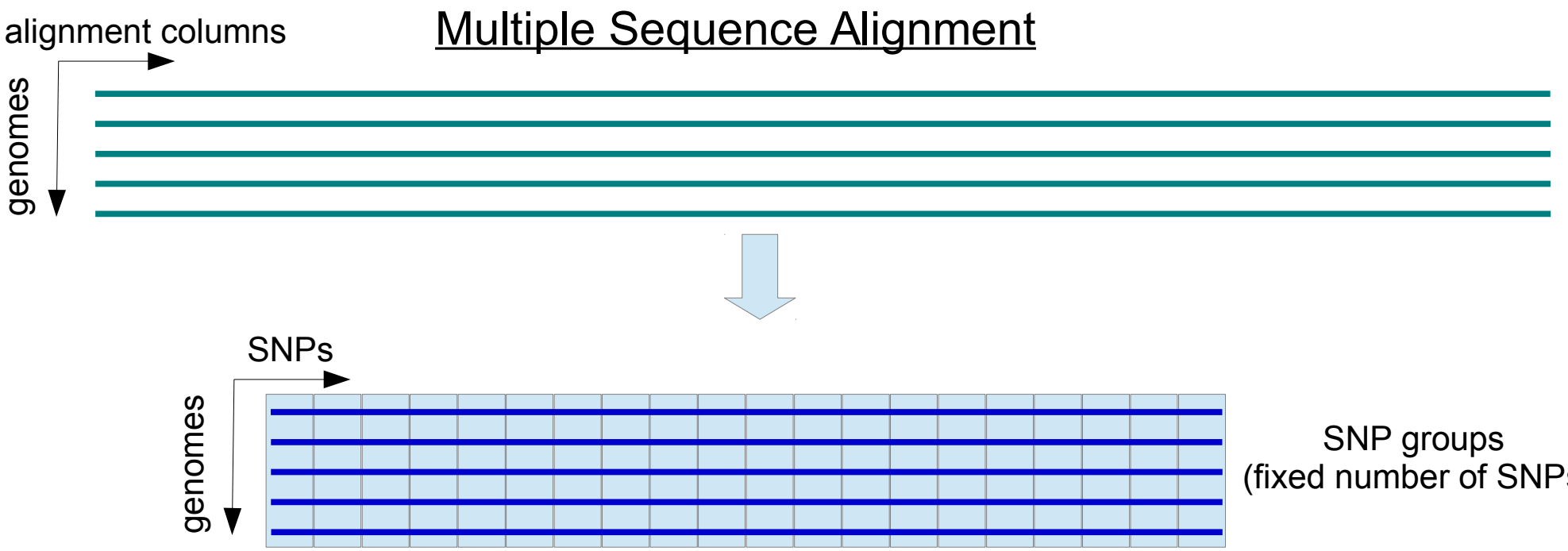
“Generic” LD approach



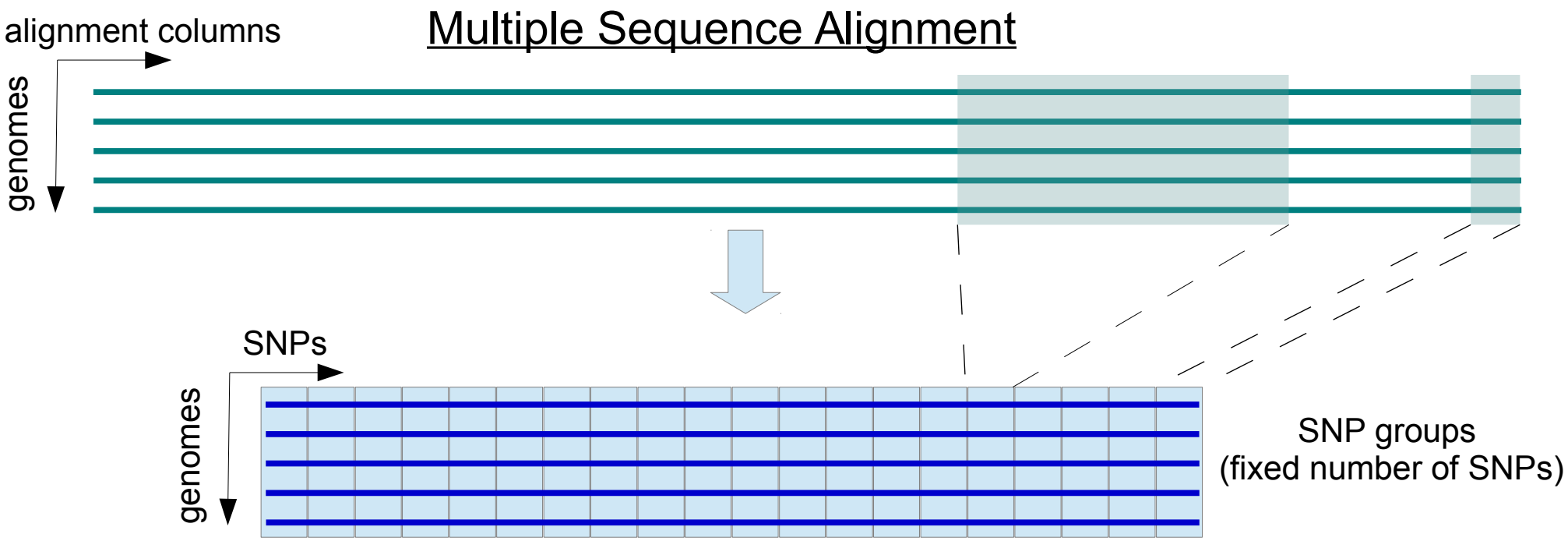
“Generic” LD approach



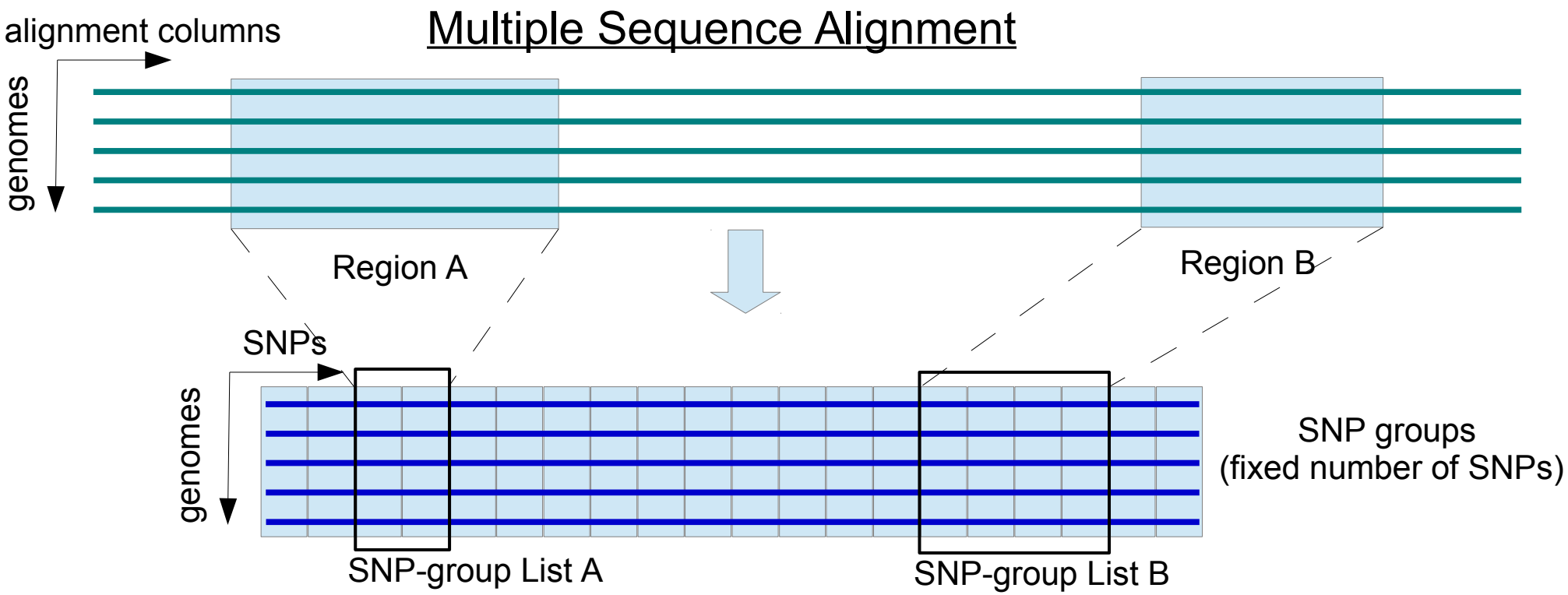
“Generic” LD approach



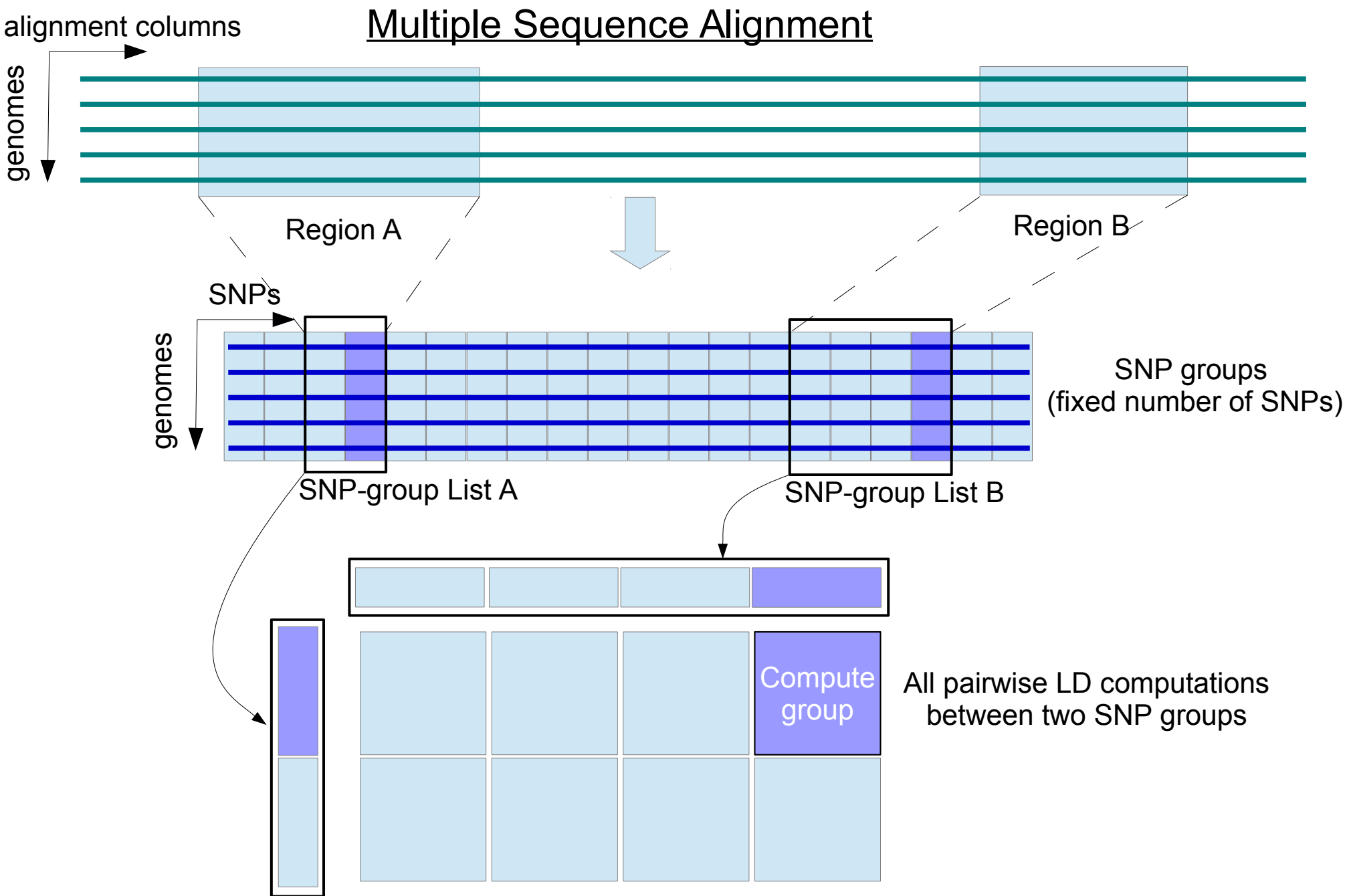
“Generic” LD approach



“Generic” LD approach



“Generic” LD approach

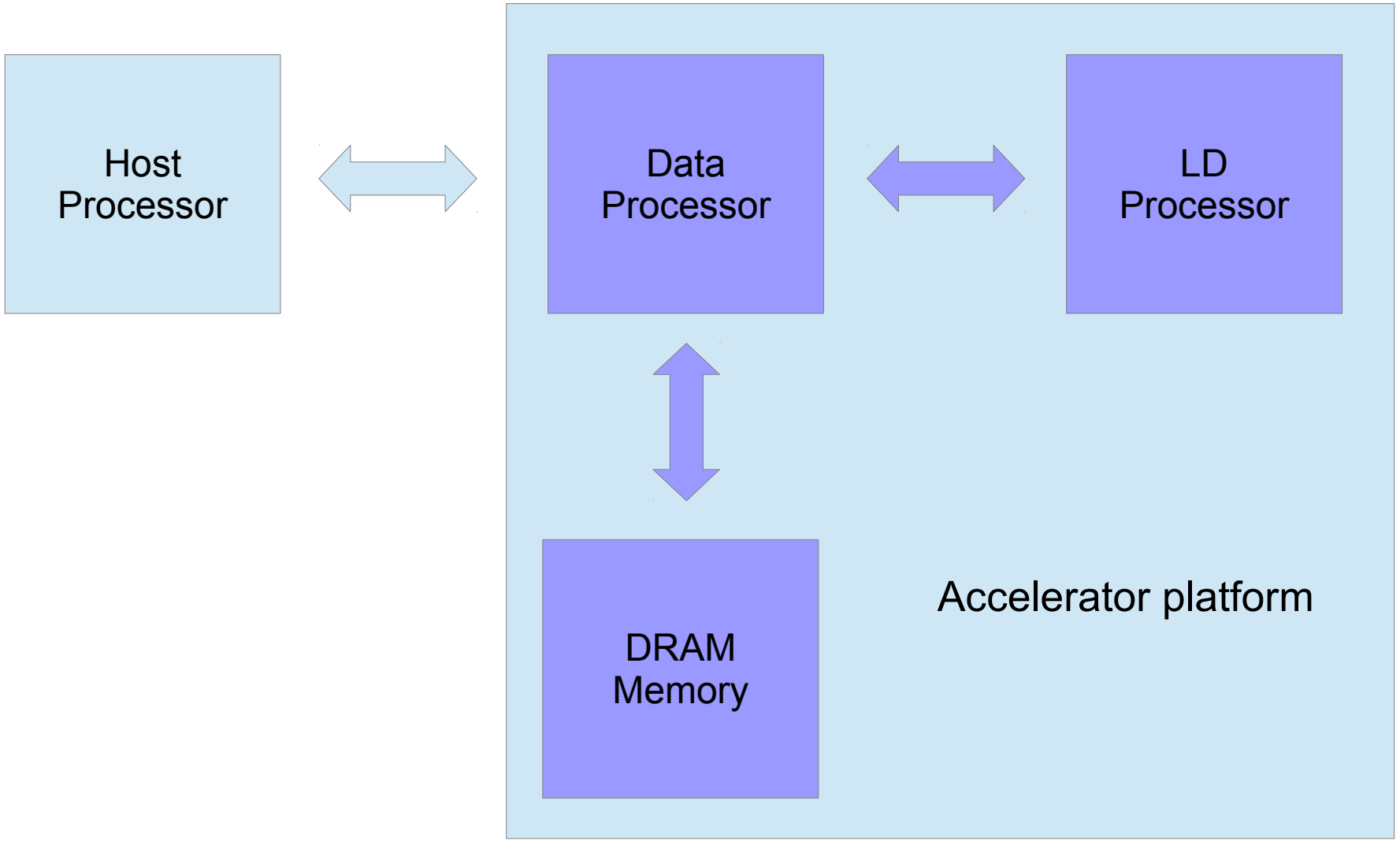


“Generic” LD approach

- No excessive memory requirements and computations when long range LD computations are conducted.
- Offloading to accelerators (FPGAs, GPUs) regardless of available memory capacity on the accelerator platform.
- Better scalability than other parallel algorithms for increasing number of cores in a processor, due to better computation-to-synchronization ratios.

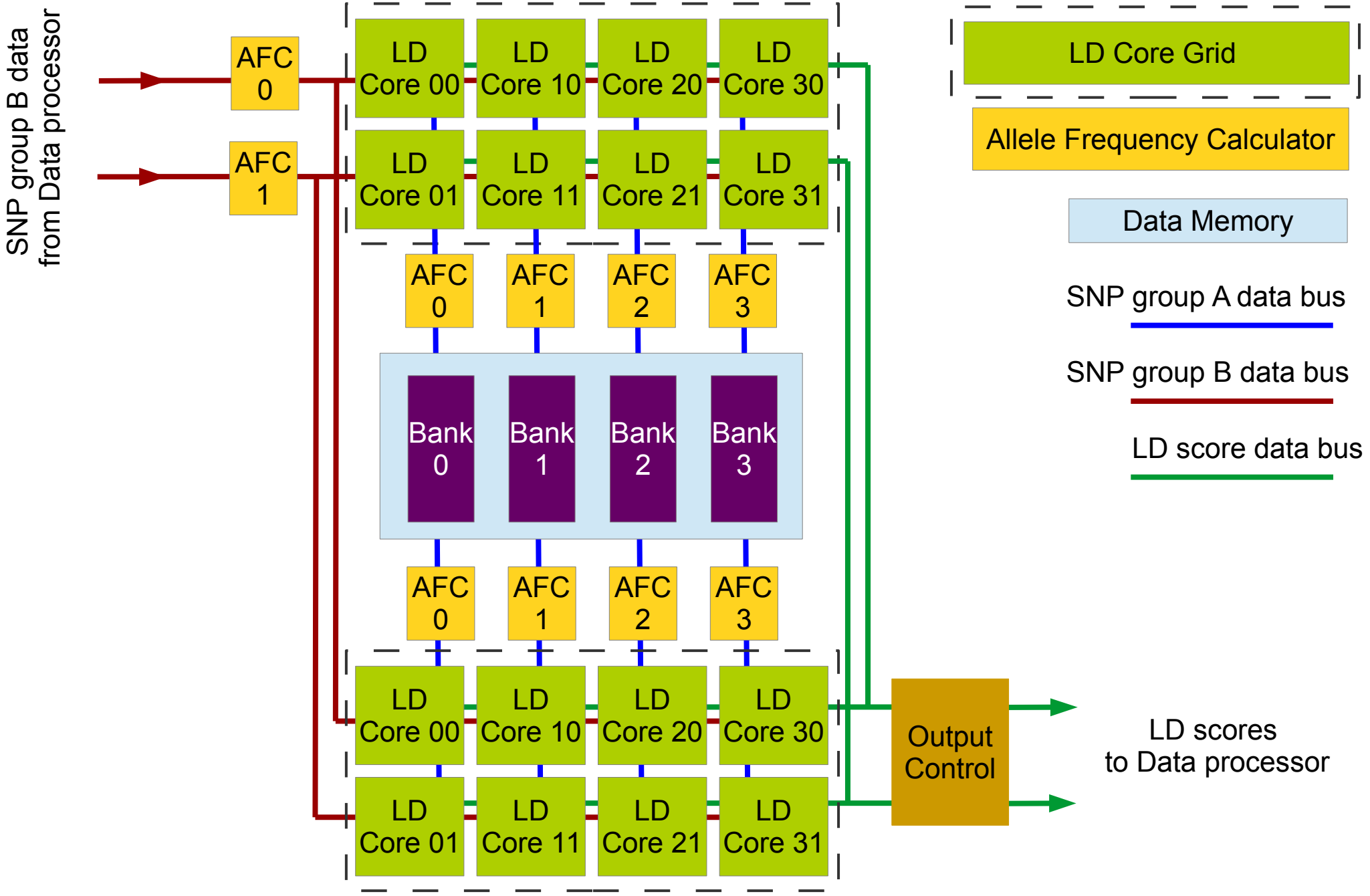
System overview

System overview

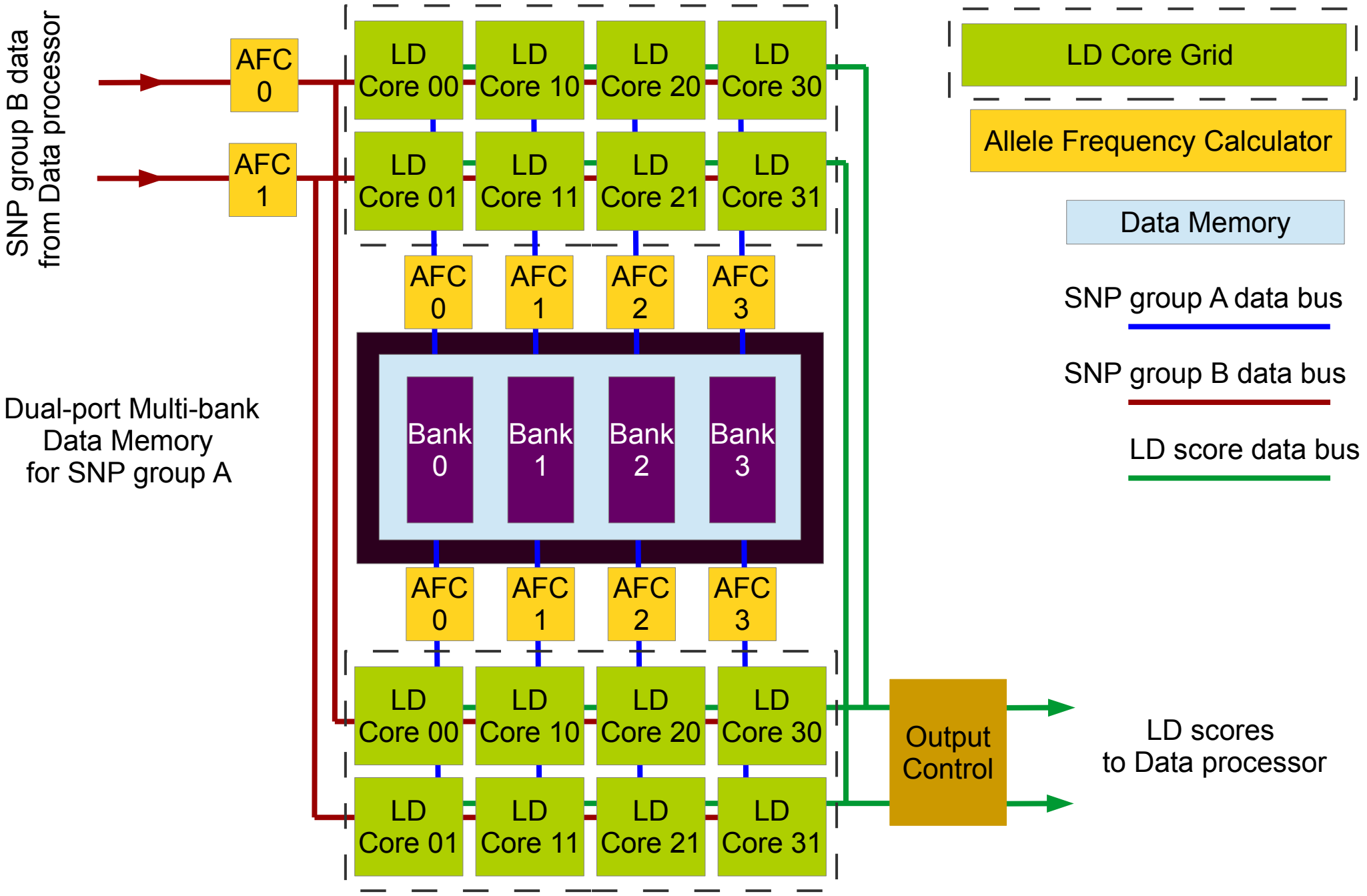


LD processor

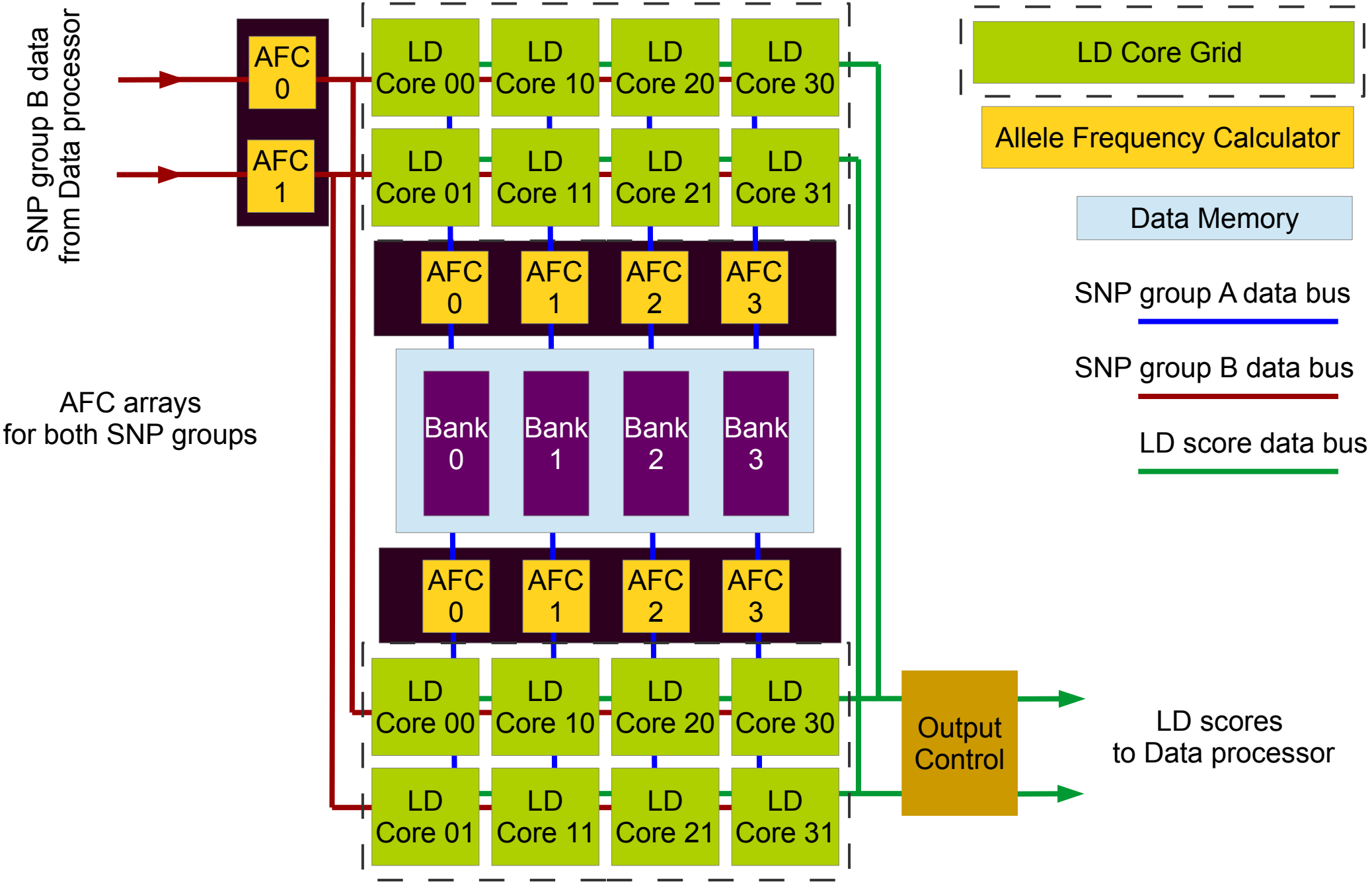
LD processor



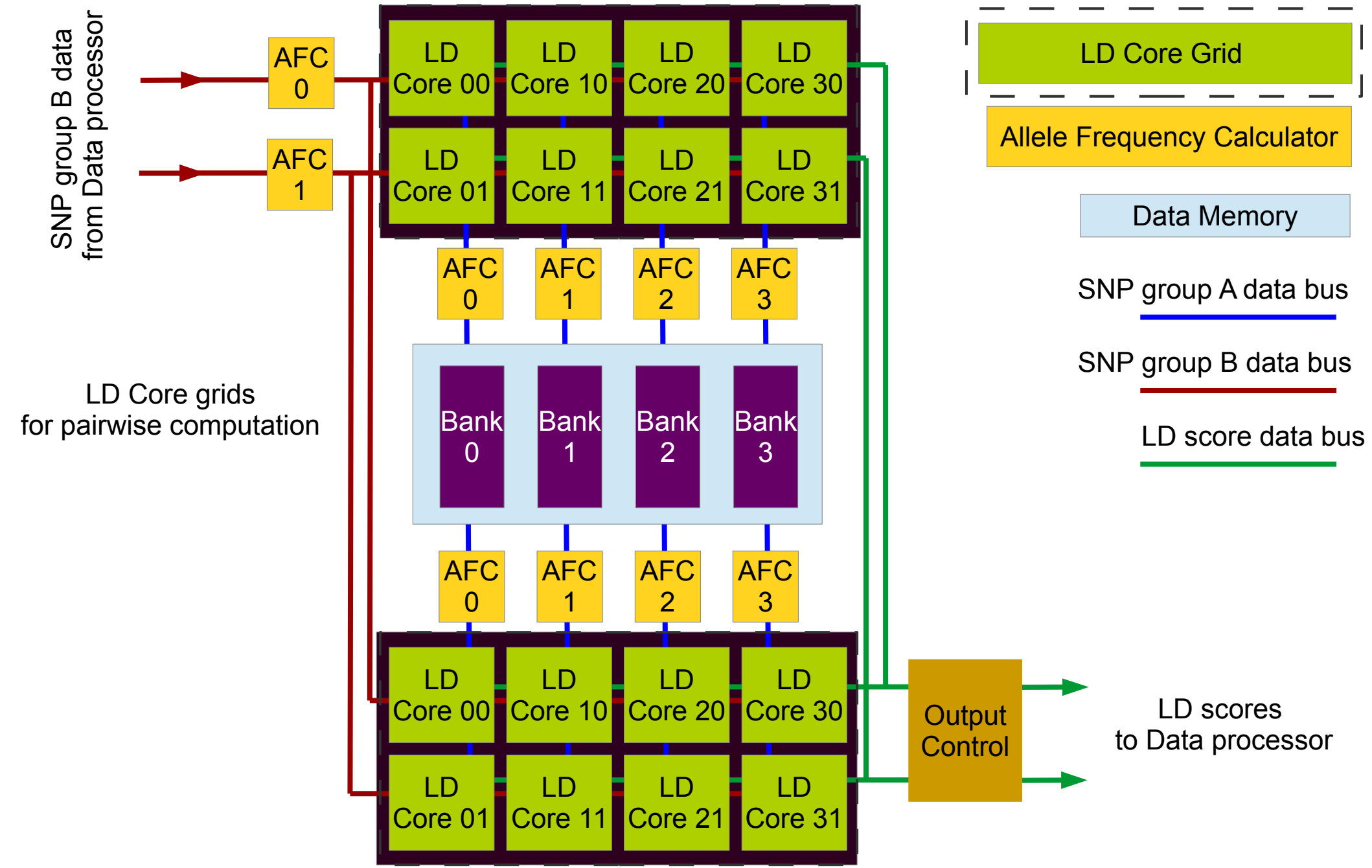
LD processor



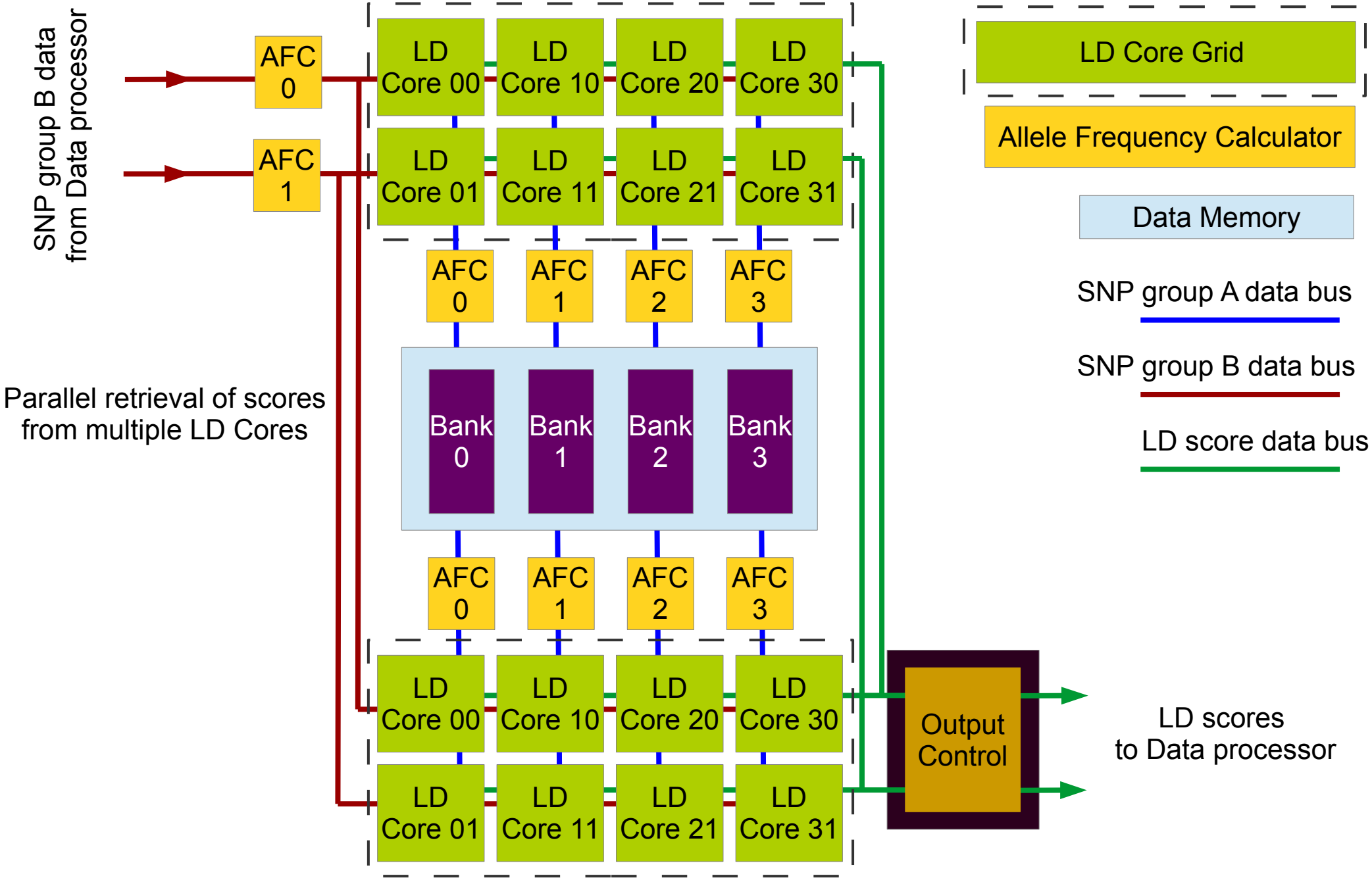
LD processor



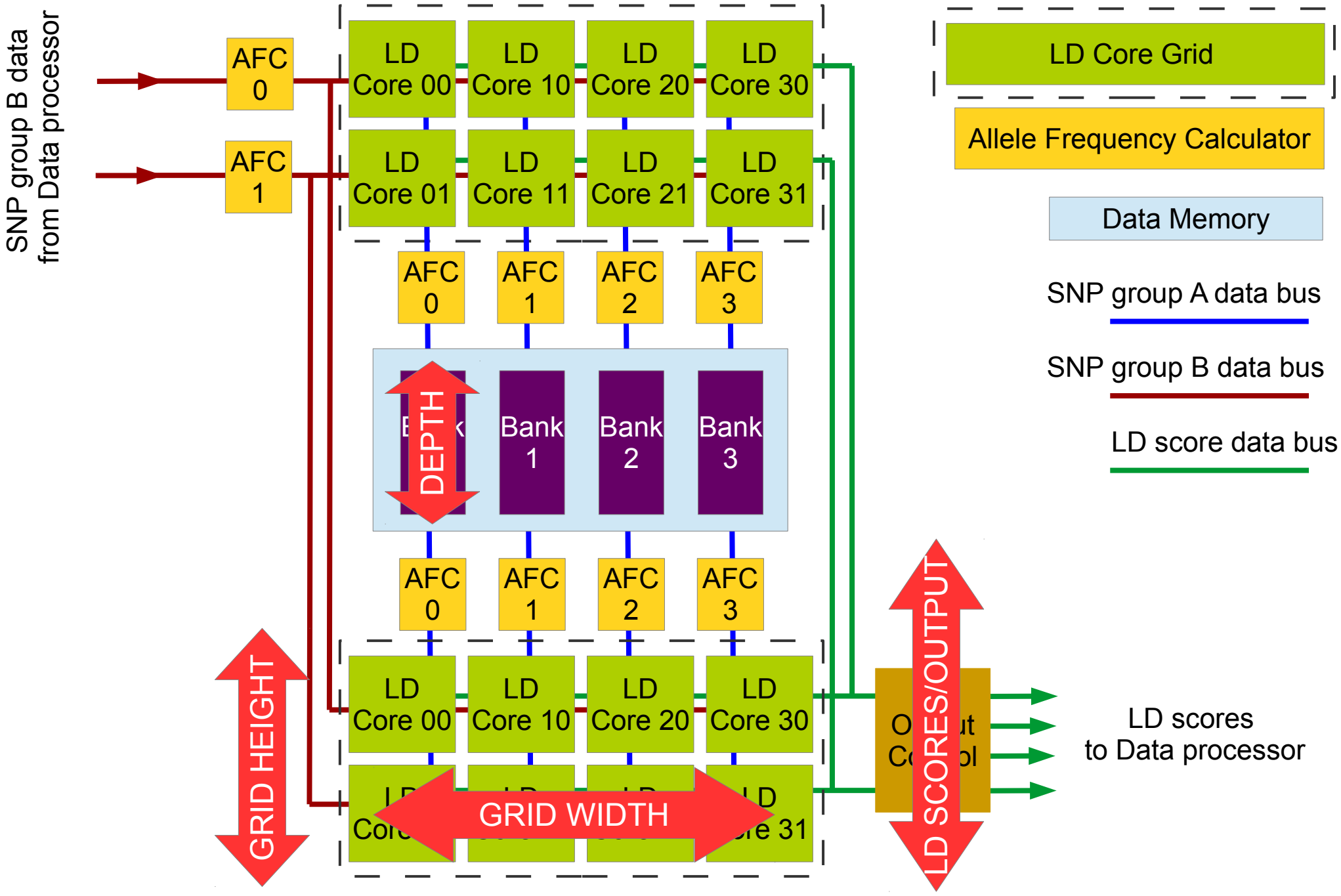
LD processor



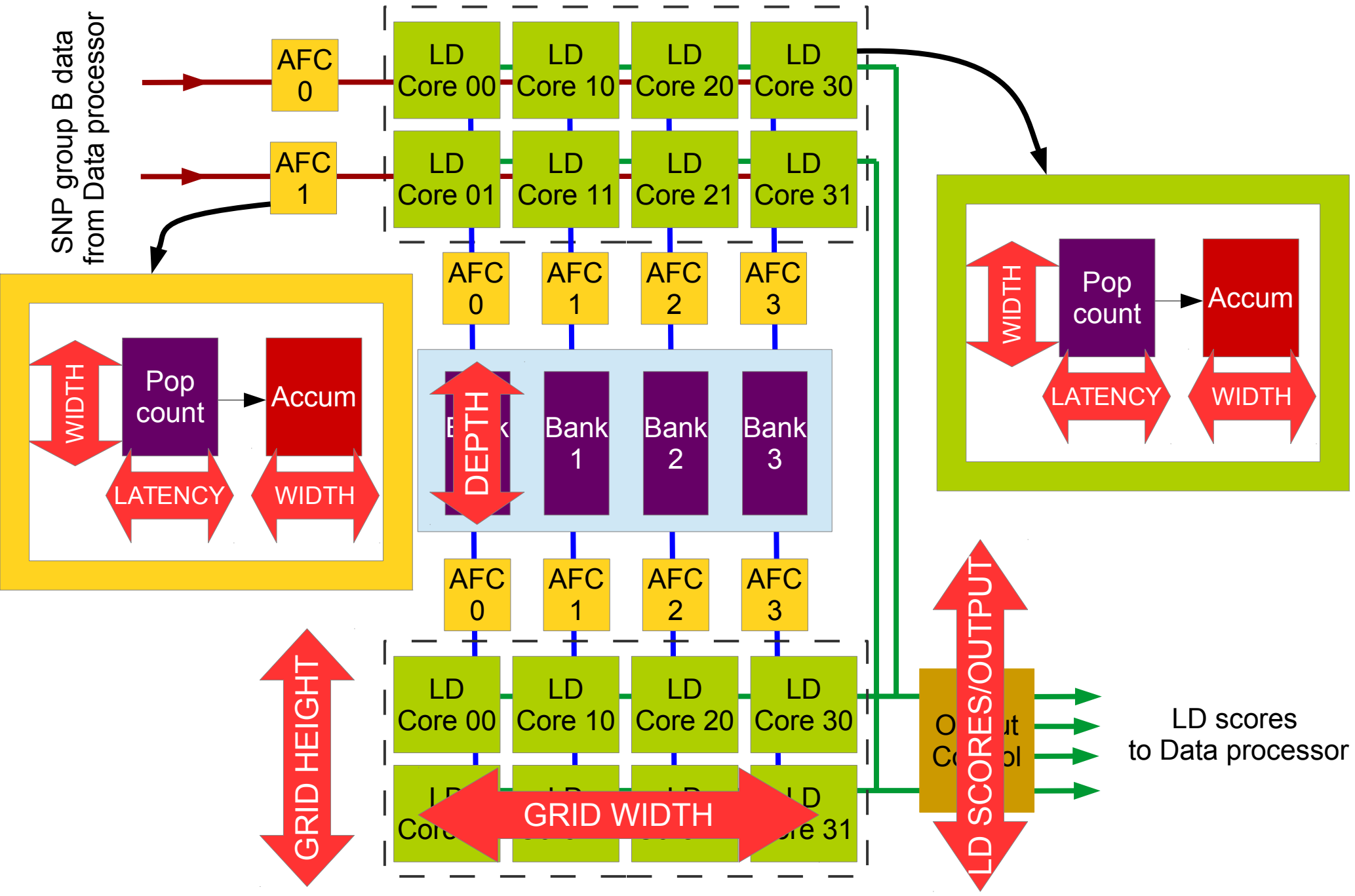
LD processor



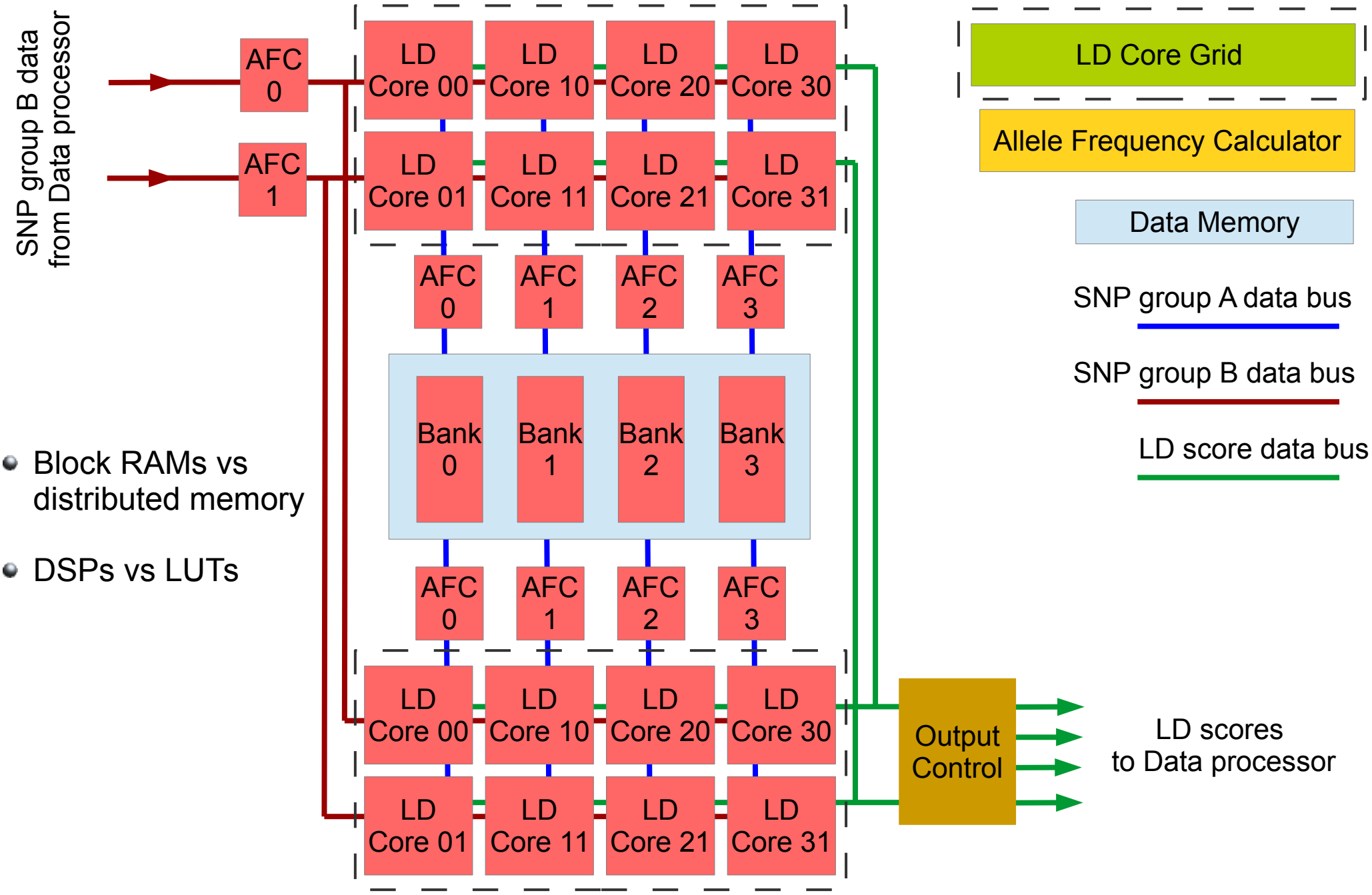
LD processor: Type S (Size) parameters



LD processor: Type S (Size) parameters



LD processor: Type R (Resources) parameters

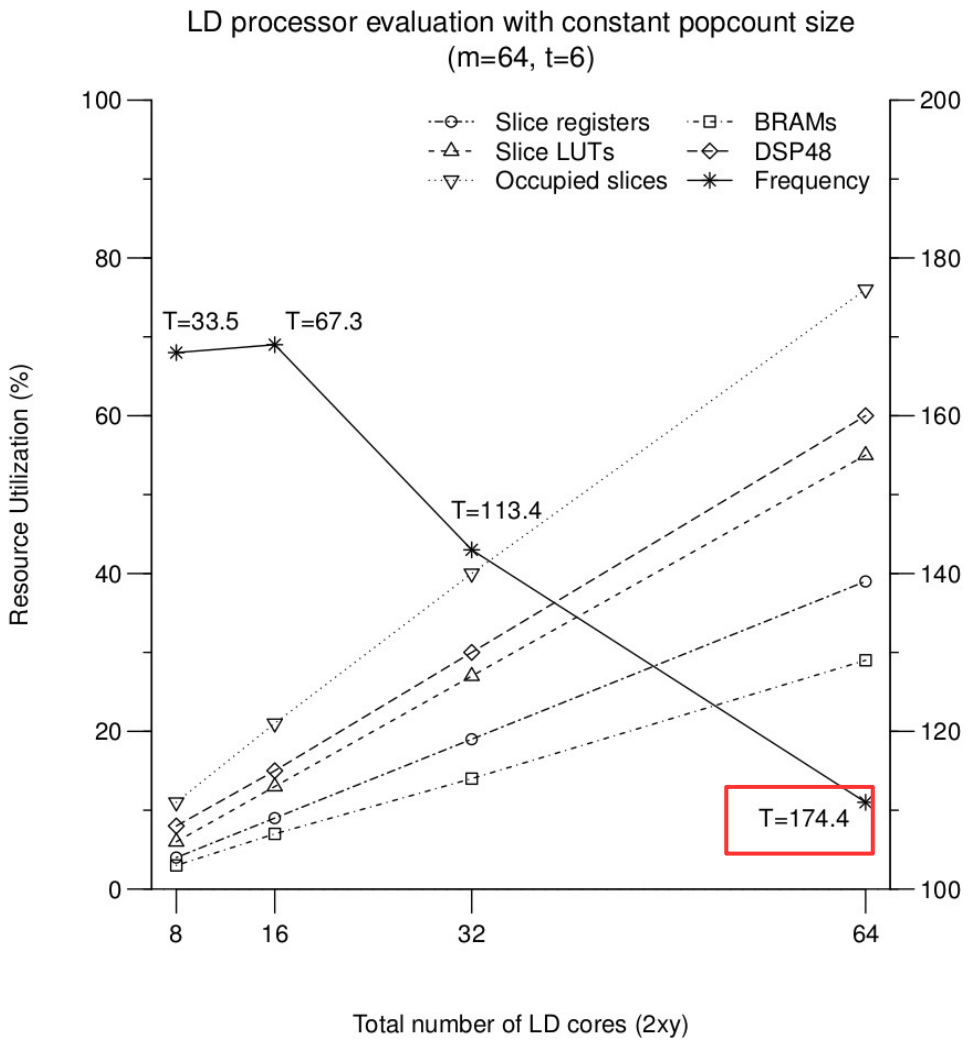


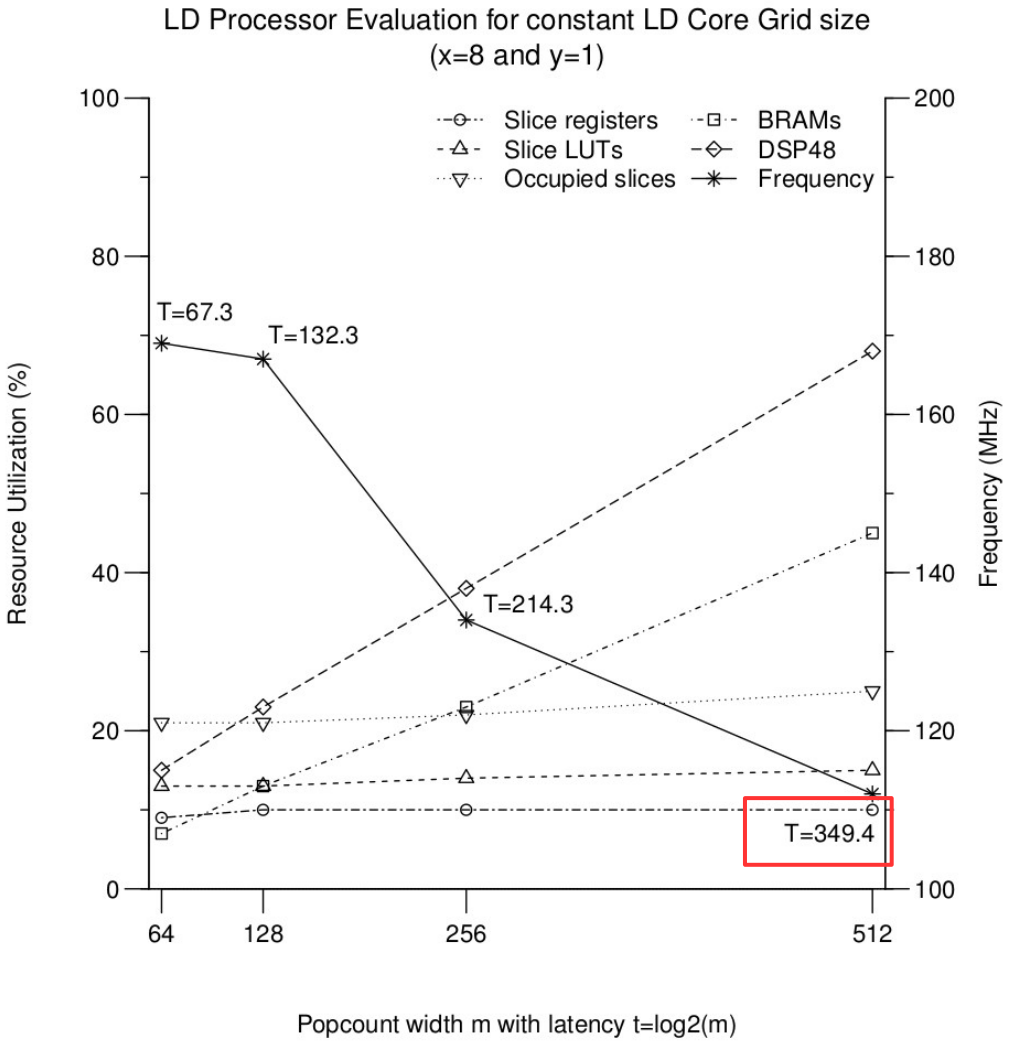
- Block RAMs vs distributed memory
- DSPs vs LUTs

Design space exploration

Design space exploration: Tuning S parameters

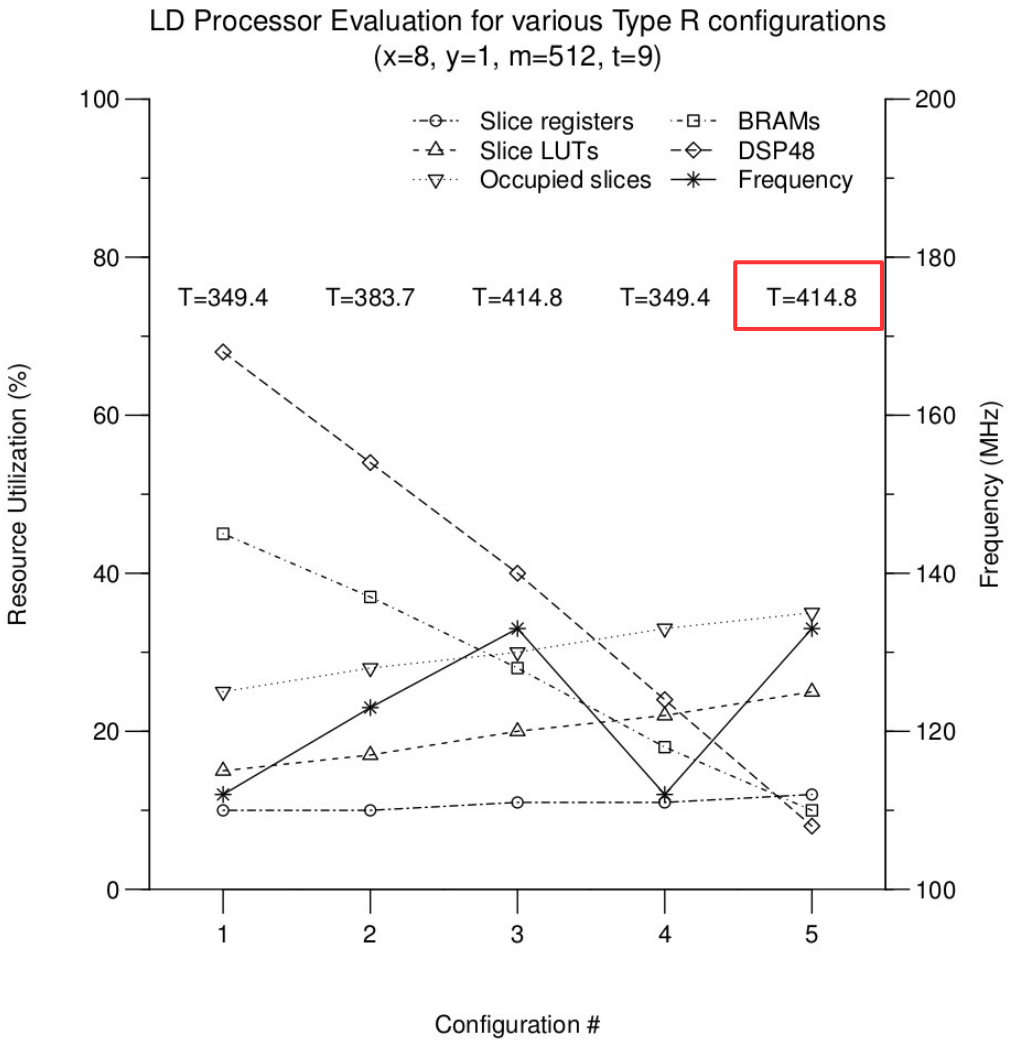
Virtex 7 VX980T-2 – post Place And Route results
Throughput (million LDs/second)
10,000 SNPs – 2,504 genomes





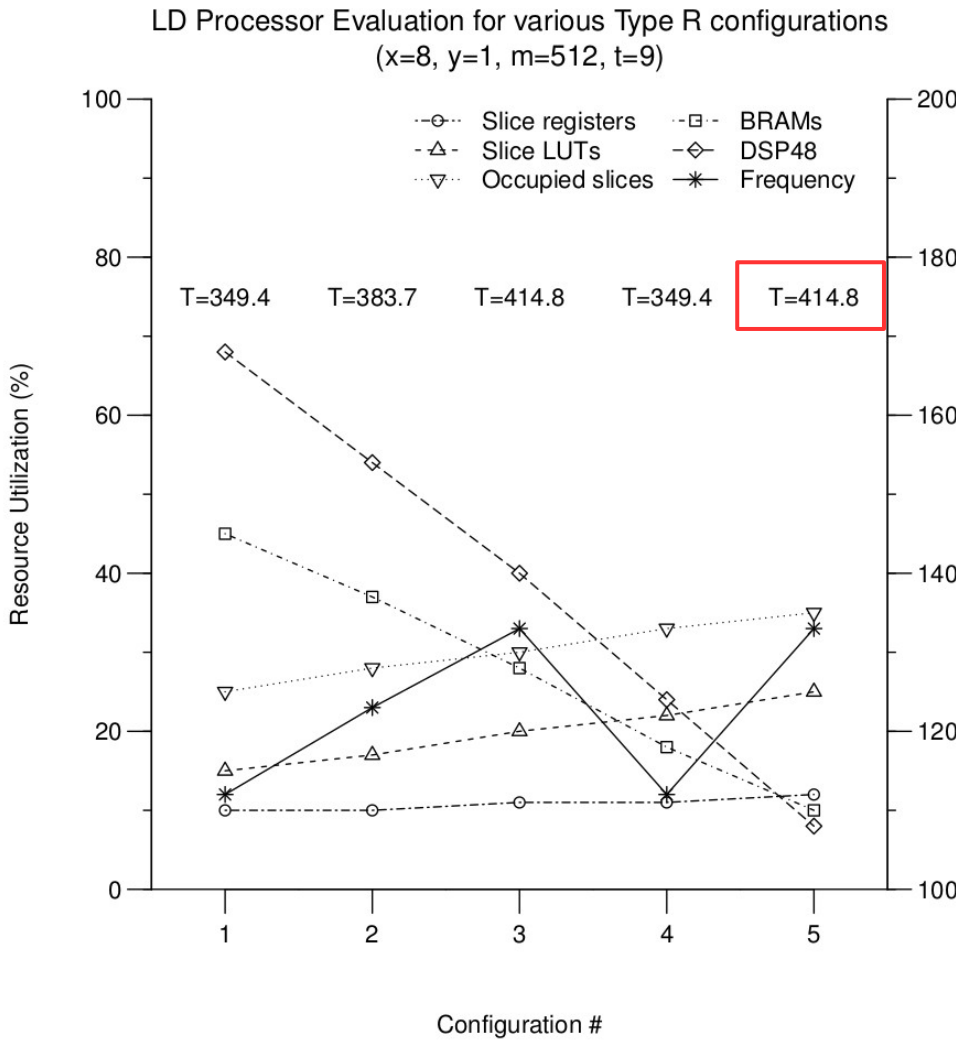
Design space exploration: Tuning R parameters

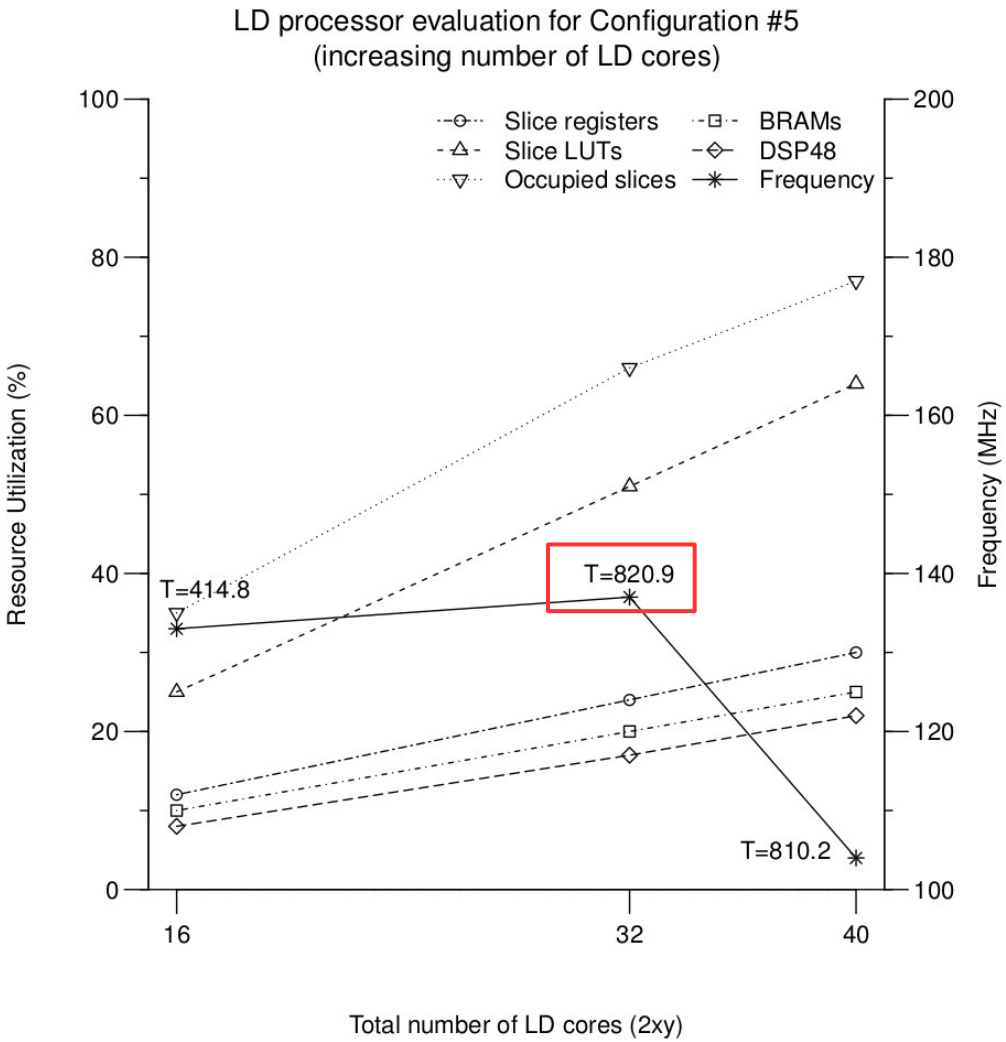
Virtex 7 VX980T-2 – post Place And Route results
Throughput (million LDs/second)
10,000 SNPs – 2,504 genomes



Design space exploration: Refining S parameters

Virtex 7 VX980T-2 – post Place And Route results
Throughput (million LDs/second)
10,000 SNPs – 2,504 genomes





Performance comparison

Performance comparison

AJHG

Cell PRESS

Volume 81, Issue 3, September 2007, Pages 559–575

Report

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell[b, a], Benjamin Neale[b, c], Kathe Todd-Brown[a], Lori Thomas[a], Manuel A.R. Ferreira[a], David Bender[b, a], Julian Maller[b, a], Pamela Sklar[b, a, a], Paul I.W. de Bakker[b, a], Mark J. Daly[b, a], Pak C. Sham[d]

Platform: Intel Xeon E5-2630 6-core processor at 2.60 GHz and 32 GBs main memory

Data: 10,000SNPs – 2,504 human genomes (real, 1000Genomes project)
10,000SNPs – 10,000 sequences (synthetic)
10,000SNPs – 100,000 sequences (synthetic)

Performance comparison

2,504 human genomes (1000Genomes project), 10,000 SNPs

Threads	PLINK 1.9 Exec. time (sec)	PLINK 1.9 mLD/sec	FPGA LD Proc. Speedup (X)
1	12.3	4.1	200.2
2	9.6	5.2	157.8
4	5.9	8.4	97.7
8	3.8	13.0	63.1
12	3.0	16.4	50.1

FPGA
820.9 mLD/sec

D.1: 10,000 sequences, D.2: 100,000 sequences, 10,000 SNPs

Threads	PLINK 1.9 Exec. time (sec)		PLINK 1.9 mLD/sec		FPGA LD Proc. Speedup (X)	
	D.1	D.2	D.1	D.2	D.1	D.2
1	41.1	389.1	1.2	0.128	171.4	159.3
2	31.4	297.6	1.6	0.168	128.5	121.4
4	19.2	180.2	2.6	0.277	79.1	73.6
8	11.3	109.4	4.4	0.456	46.8	44.7
12	9.9	88.3	5.0	0.566	41.1	36.0

FPGA
D.1: 205.7 mLD/sec
D.2: 20.4 mLD/sec

Conclusion

Conclusion

- Lack of a vectorized population count operator in processors does not permit the implementation of a high performance microkernel
- FPGAs allow very wide and deep bit-counting pipelines, which is a key for performance as more genomes are sequenced
- Hardware generation software to explore the design space and get high performance for a particular dataset size

Thank you