Efficient Binarization for Historical Document Analysis

Florian Westphal Håkan Grahn Niklas Lavesson

Blekinge Institute of Technology
Karlskrona, Sweden

flw@bth.se

2016-02-02

Outline

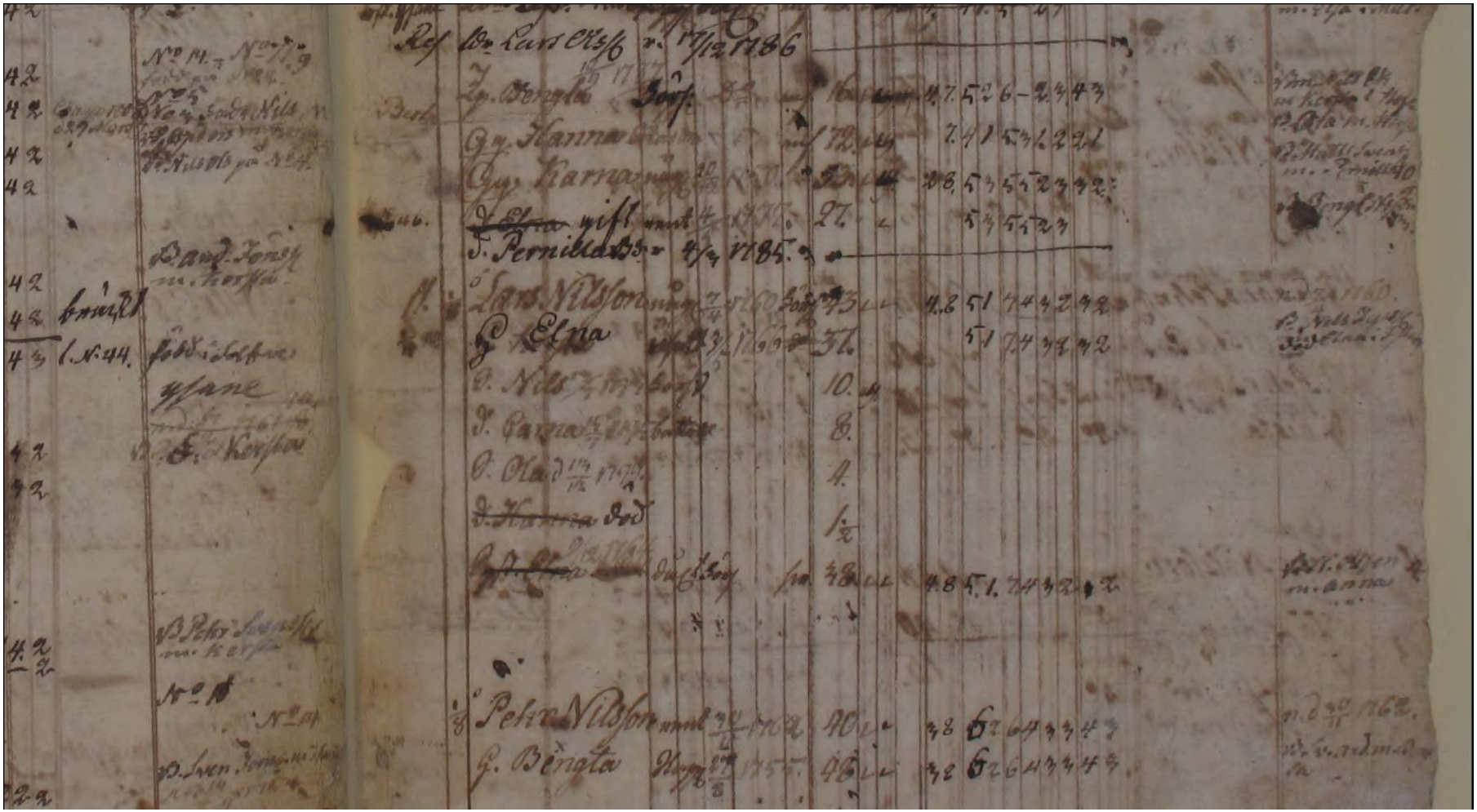
1. Document Readability
2. Howe's Binarization Algorithm
3. Heterogenous Computing
4. Binarization Pipeline

BTH & ArkivDigital

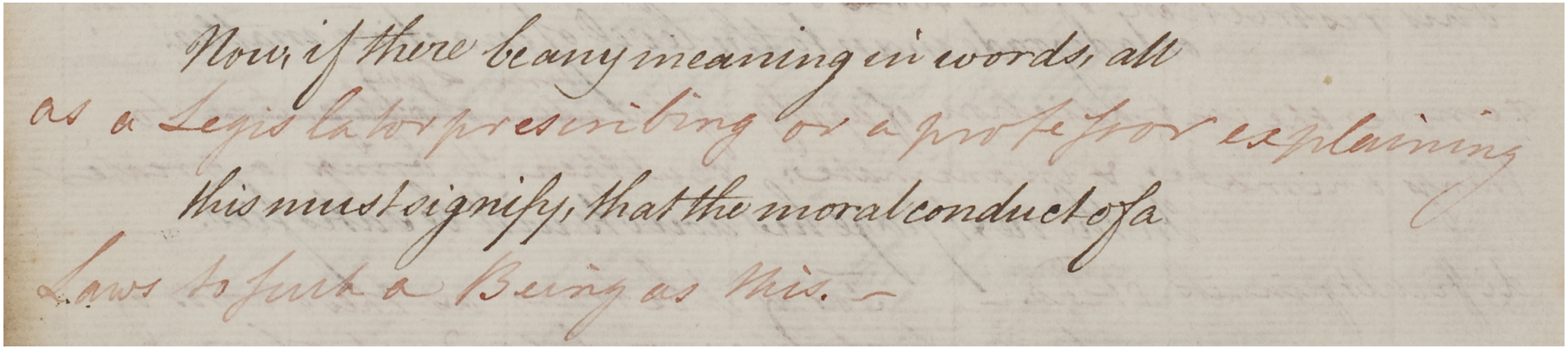
- Swedish university, established in 1989
- Over 6000 registered students
- BigData@BTH

- Swedish company, established in 2004
- Provides access to almost 60 million images
- Church books, court records, military records, census records, ...

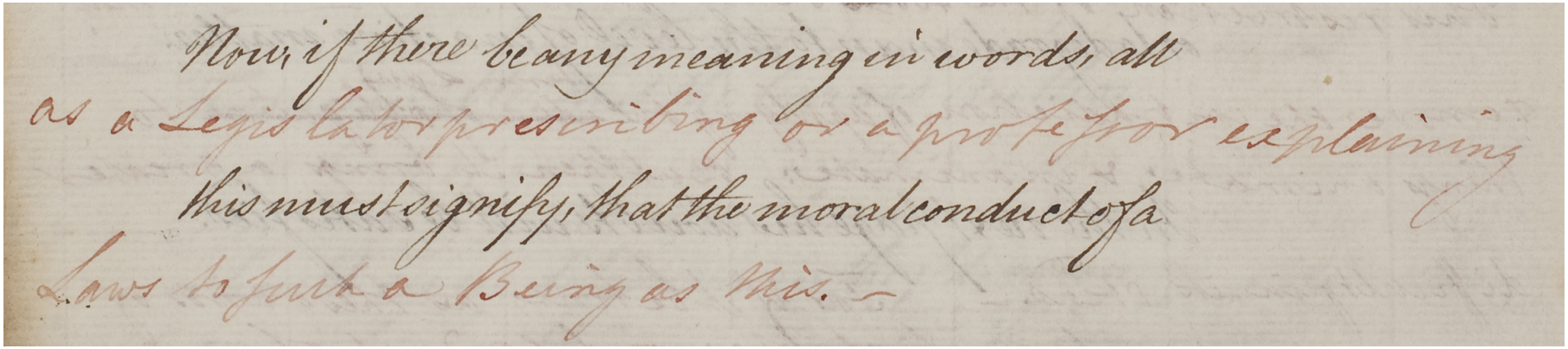
Document Readability



Approach

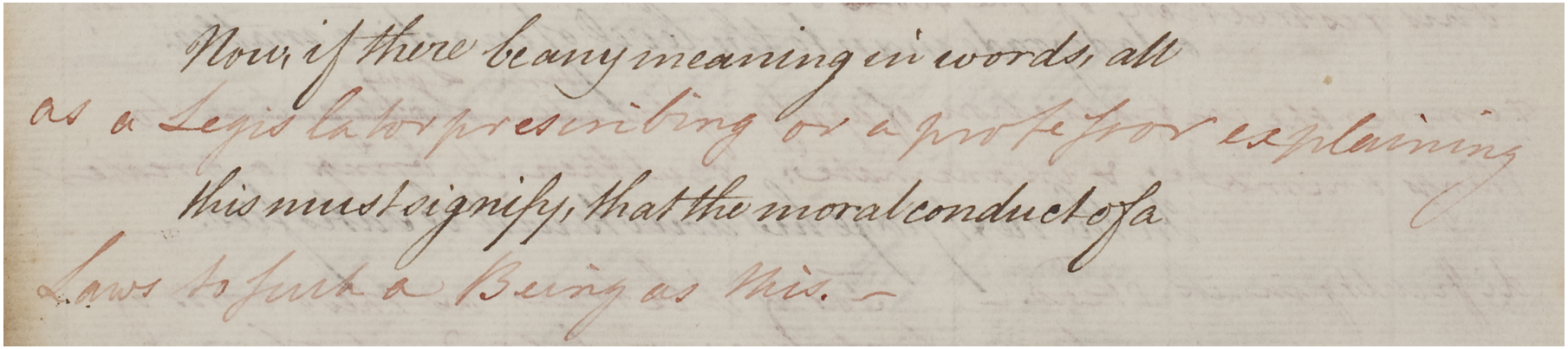

Now if there be any meaning in words, all
as a Legislator prescribing or a professor explaining
this must signify, that the moral conduct of a
Laws to suit a Being as this. –

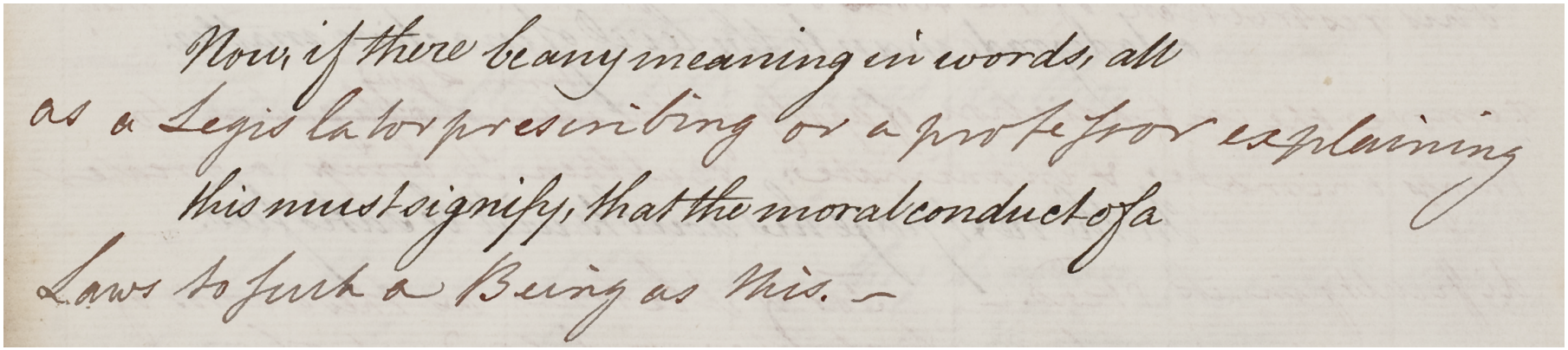
Approach



Now if there be any meaning in words, all
as a Legislator prescribing or a professor explaining
this must signify, that the moral conduct of a
Laws to such a Being as this. –

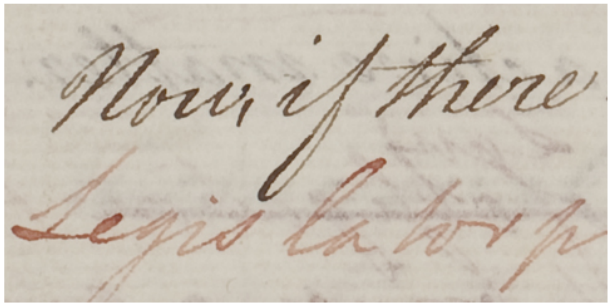
Approach


Now if there be any meaning in words, all
as a Legislator prescribing or a professor explaining
this must signify, that the moral conduct of a
Laws to such a Being as this. –

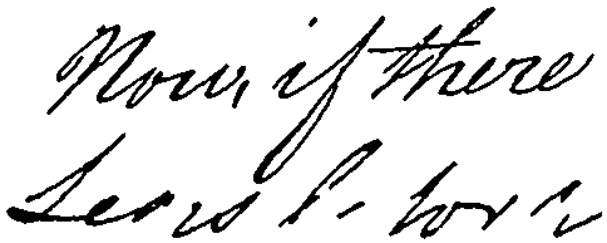

Now if there be any meaning in words, all
as a Legislator prescribing or a professor explaining
this must signify, that the moral conduct of a
Laws to such a Being as this. –

Approach - Demo

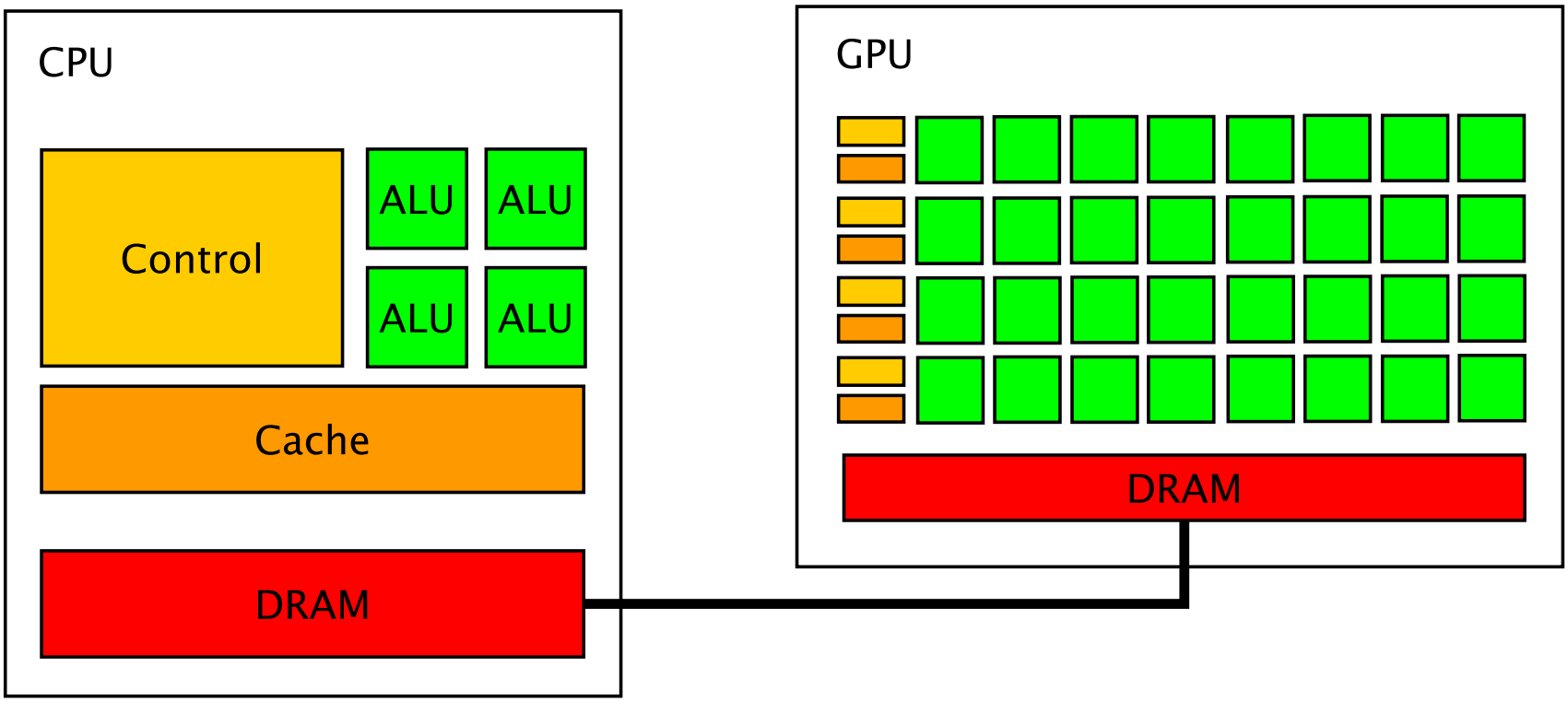
Howe's Binarization Algorithm



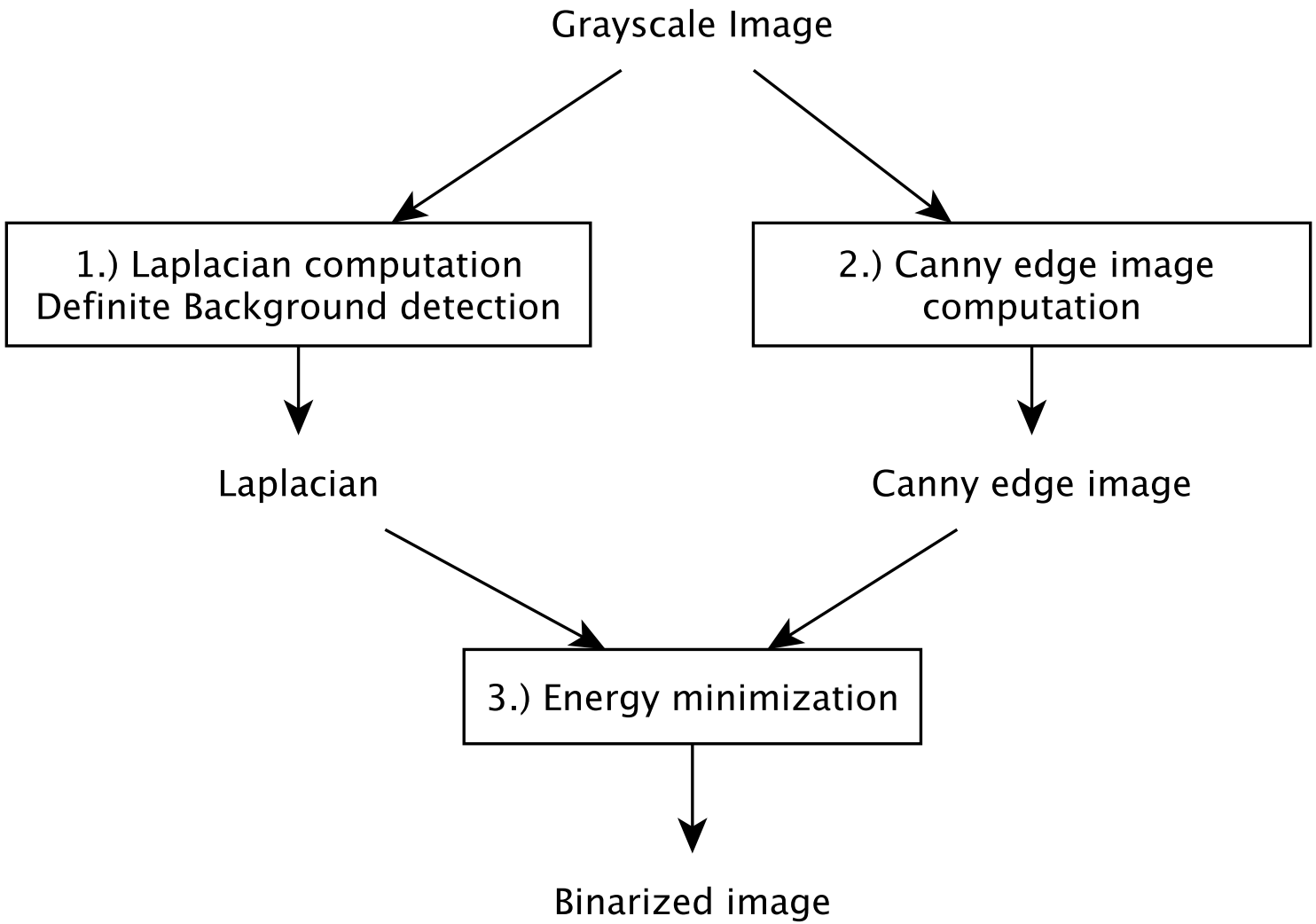
Howe's Binarization Algorithm (Cont.)



Heterogenous Computing



Binarization Pipeline

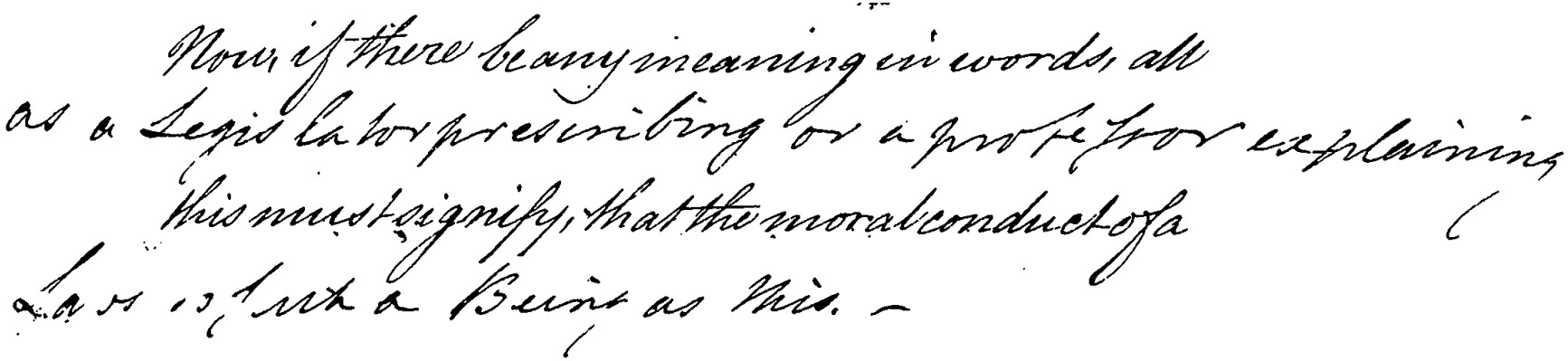


Binarization Pipeline (Cont.)

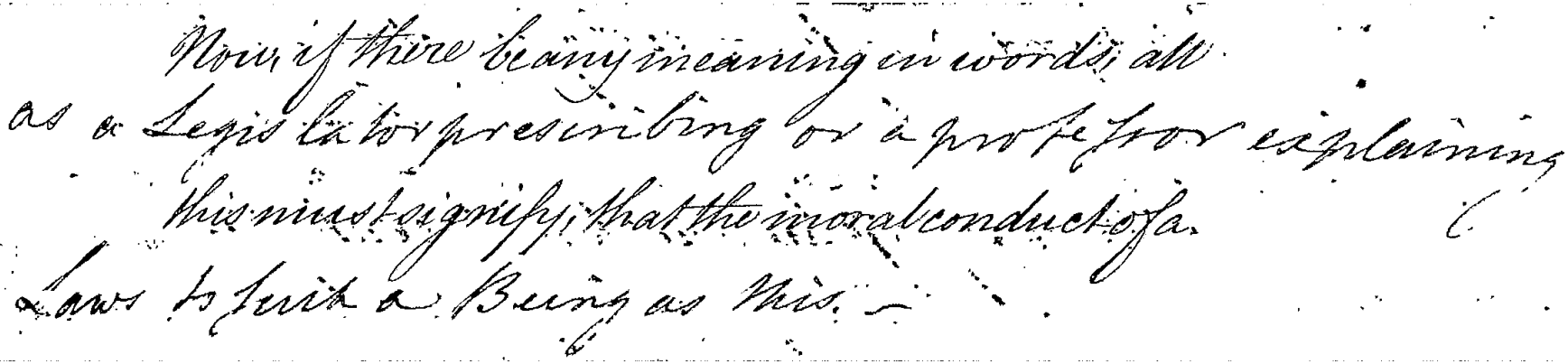
	I	II	III	IV	V	VI	VII	VIII
1	CPU	GPU	CPU	GPU	CPU	GPU	CPU	GPU
2	CPU	CPU	GPU	GPU	CPU	CPU	GPU	GPU
3	CPU	CPU	CPU	CPU	GPU	GPU	GPU	GPU

Preliminary Results

Reference Implementation

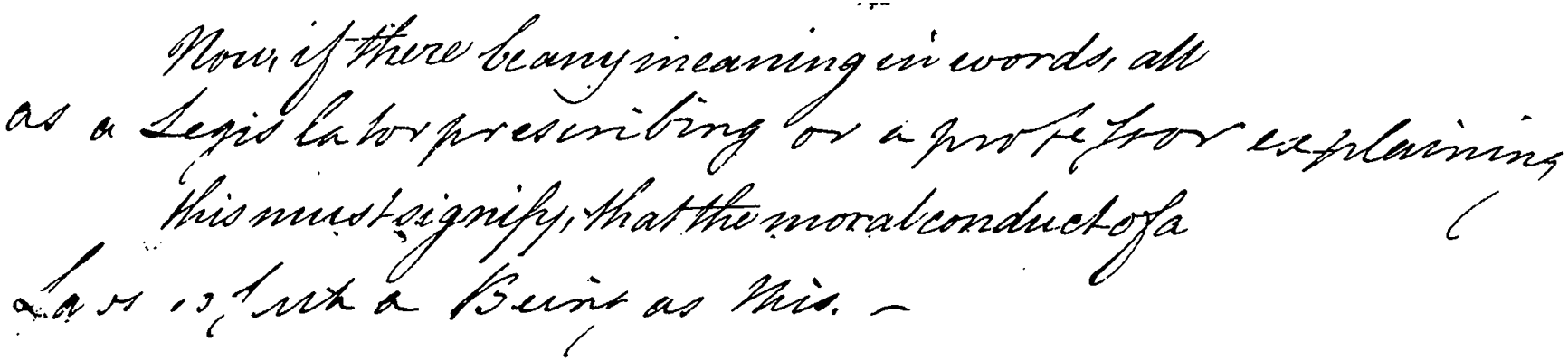


Configuration IV

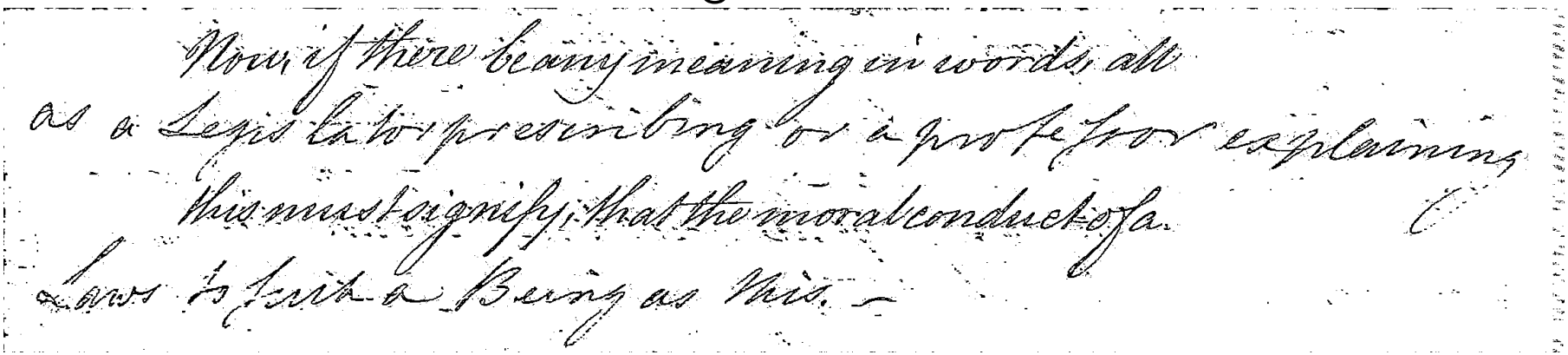


Preliminary Results (Cont.)

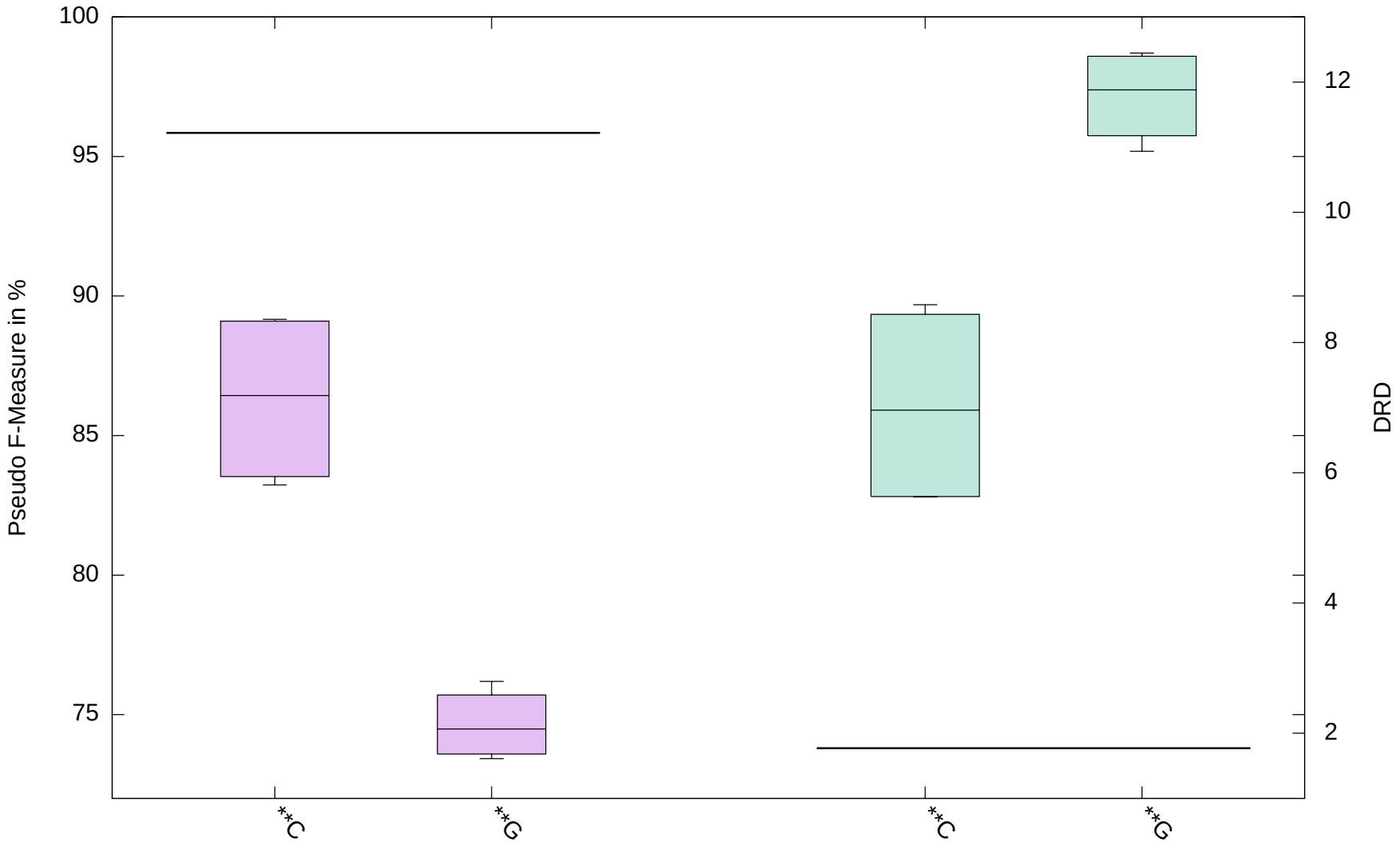
Reference Implementation



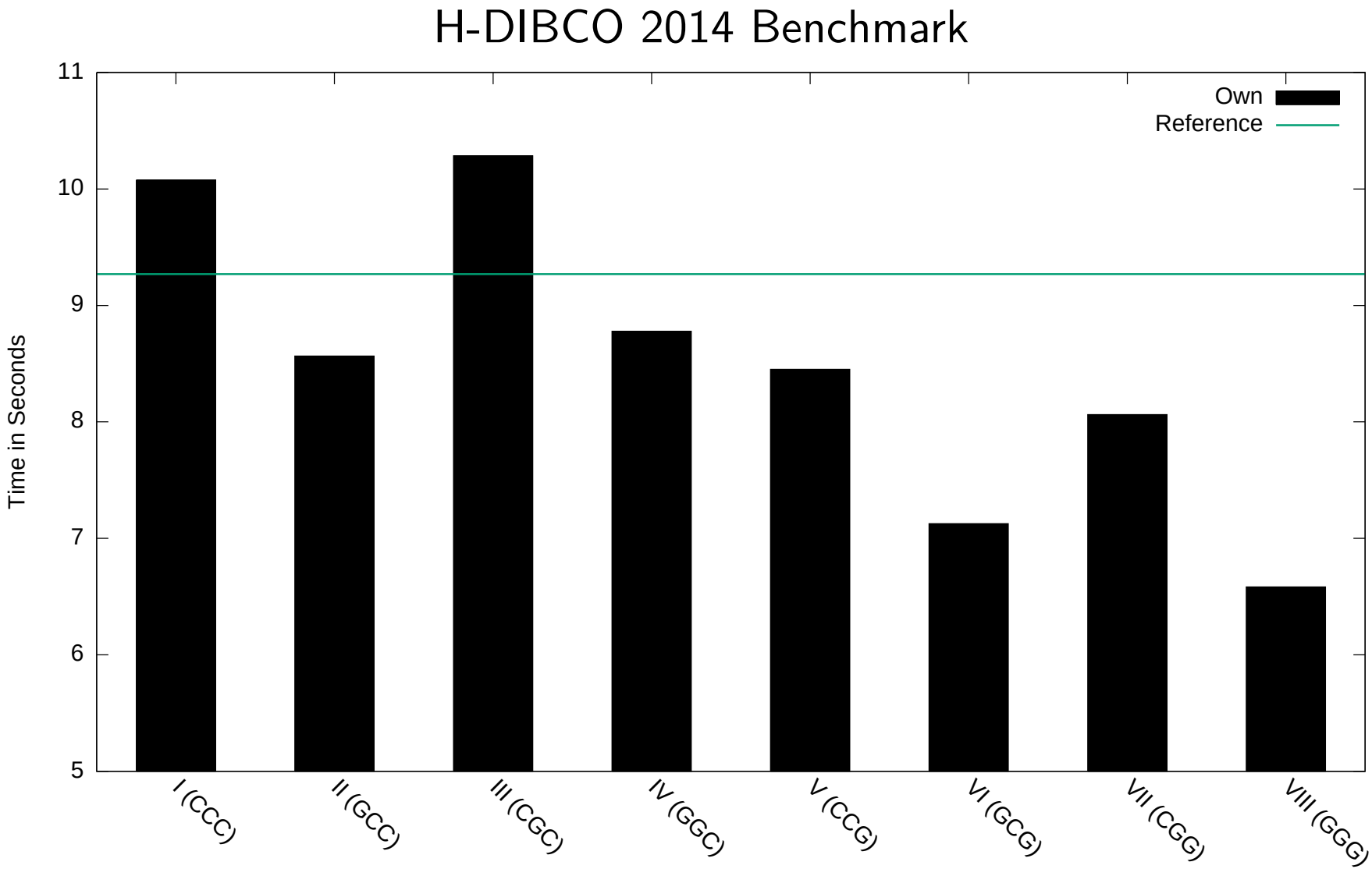
Configuration VIII



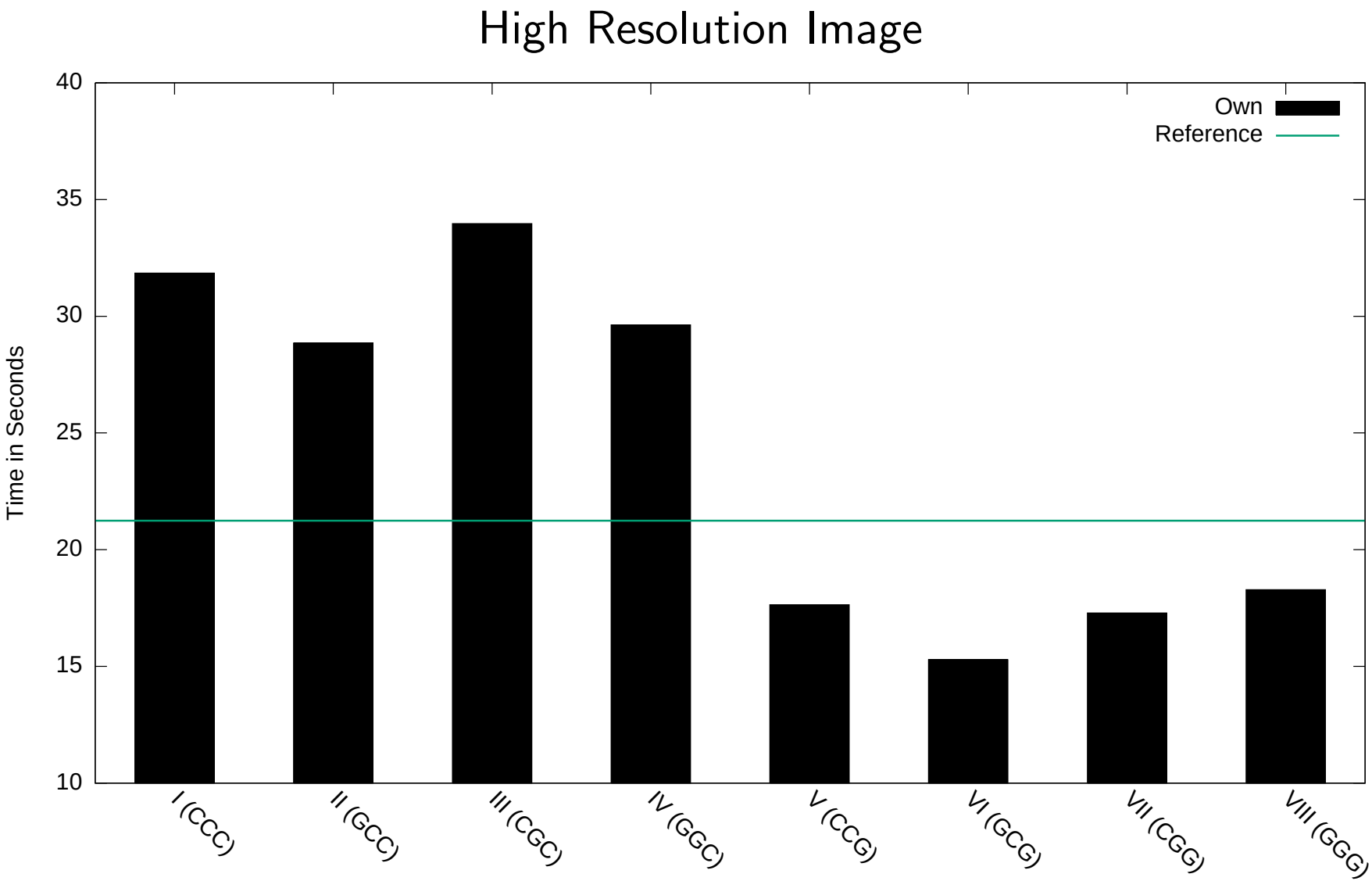
Preliminary Results - Binarization Performance



Preliminary Results - Time



Preliminary Results - Time (cont.)



Preliminary Results - Time per Step

Time taken in each binarization step for the used high resolution image.

	1	2	3
CPU	2.27 s	0.17 s	28.76 s
GPU	0.39 s	0.11 s	14.54 s

Next Steps

- Revision of the implementation
- Implementation of the binarization pipeline

Acknowledgements

We would like to thank ArkivDigital for providing us with access to their image database.

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden.