12th IEEE International Conference on eScience
(eScience 2016)

October 24-26, 2016 – Baltimore, Maryland, USA

An *n*-gram cache for large-scale parallel extraction of multiword relevant expressions with LocalMaxs

Carlos Gonçalves[1,2]; Joaquim F. Silva[2]; José C. Cunha[2]

[1] Instituto Superior de Engenharia de Lisboa, Portugal
[2] NOVA LINCS, Faculdade Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

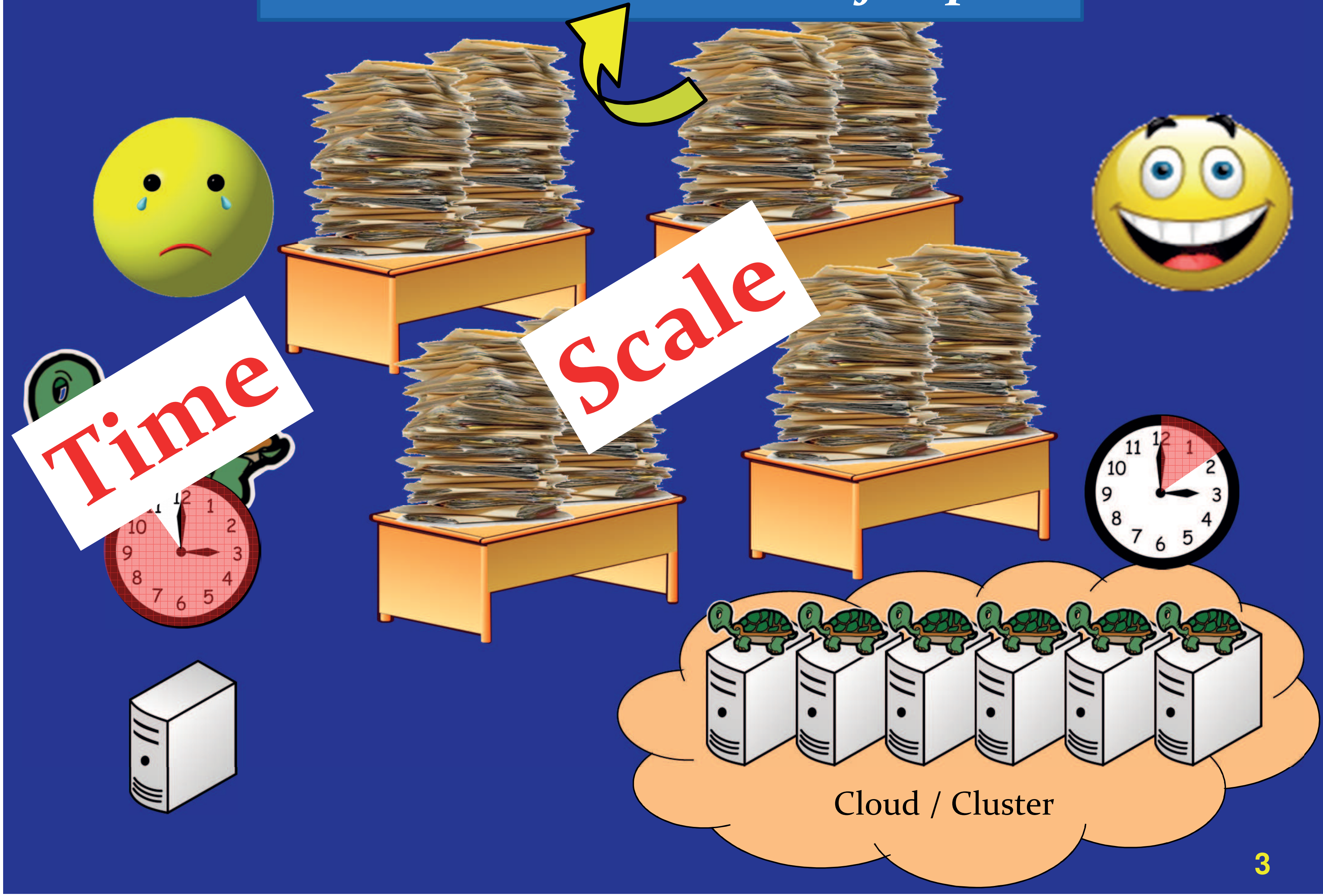
Presented by Carlos Gonçalves (cgoncalves@deetc.isel.pt)

Acknowledgments: FCT MCTES, NOVA LINCS UID/CEC/04516/2013; Lunacloud (http://www.lunacloud.com); ISEL "AzoresCloud"

Agenda

- **Objective**
- **LocalMaxs**
- **Distributed Architecture**
- ***n*-grams Statistical Distribution**
- ***n*-gram Cache System**
- **Experimental Results**
- **Conclusions and Further Work**

Statistical Extraction of Topics



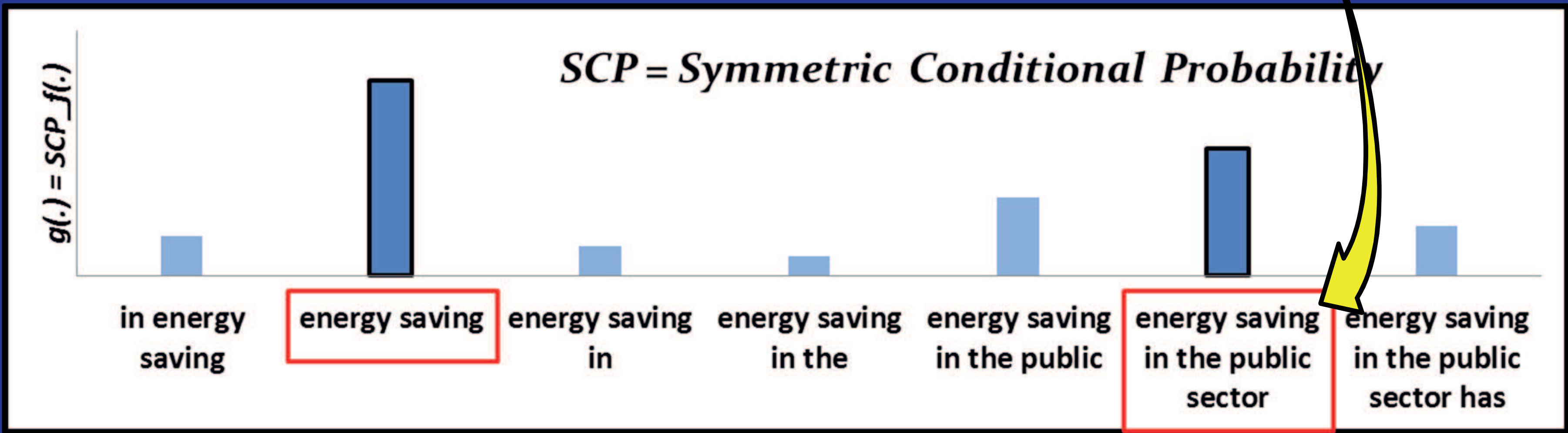
LocalMaxs: Statistical Extraction

Relevance ≡ Strong internal co-occurrence (glue) of words

$$SCP_f(w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \dots w_i) * p(w_{i+1} \dots w_n)}$$

n-gram

Detecting local maxima of the glue



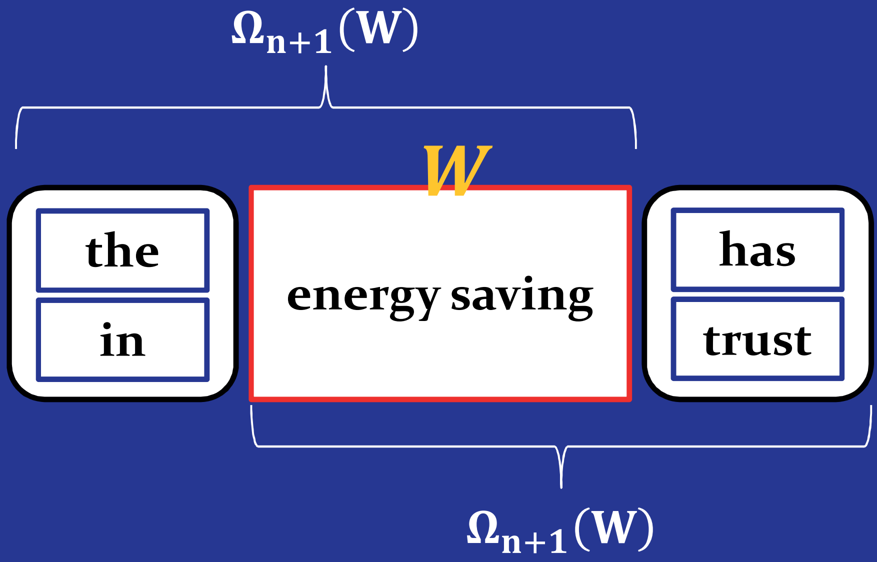
J. F. da Silva and G. P. Lopes. "A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units"

LocalMaxs: Relevance Criterion

Given a *corpus*, LocalMaxs algorithms extracts a set of multiword relevant expressions

A) Bigrams case: length(W) = 2

Comparing to adjacent enclosing *(n+1)*-grams → *3*-grams



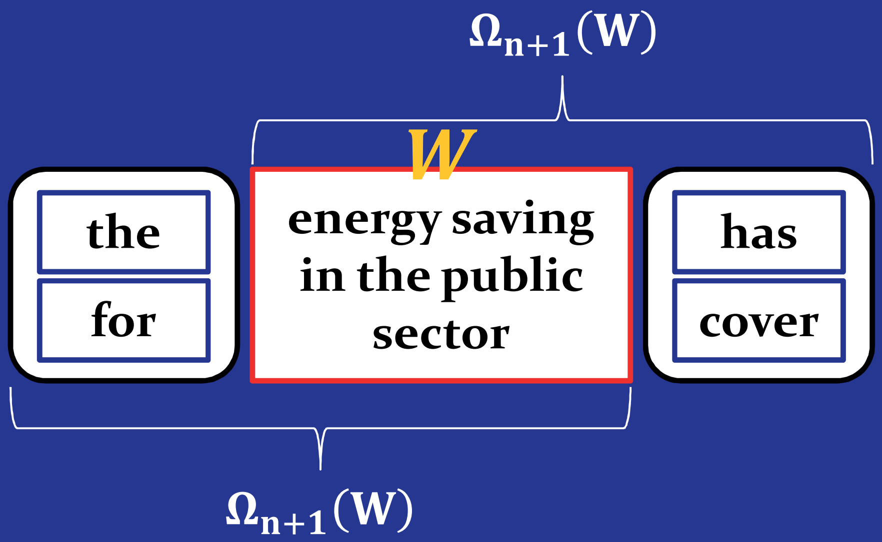
for all glue y $\in \Omega_{n+1}(W)$

$SCP(W) > y$ → **W is Relevant !**

LocalMaxs: Relevance Criterion

B) Higher *n*-grams: length(W) > 2

Comparing to the adjacent enclosing (*n+1*)-grams



LocalMaxs: Relevance Criterion

B) Higher *n*-grams: length(W) > 2

Comparing to the adjacent enclosed *(n-1)*-grams

W

energy saving in the public sector

$$\underbrace{\textbf{energy saving in the public}}_{\Omega_{n-1}(W)}$$

LocalMaxs: Relevance Criterion

B) Higher *n*-grams: length(W) > 2

Comparing to the adjacent enclosed *(n-1)*-grams

W

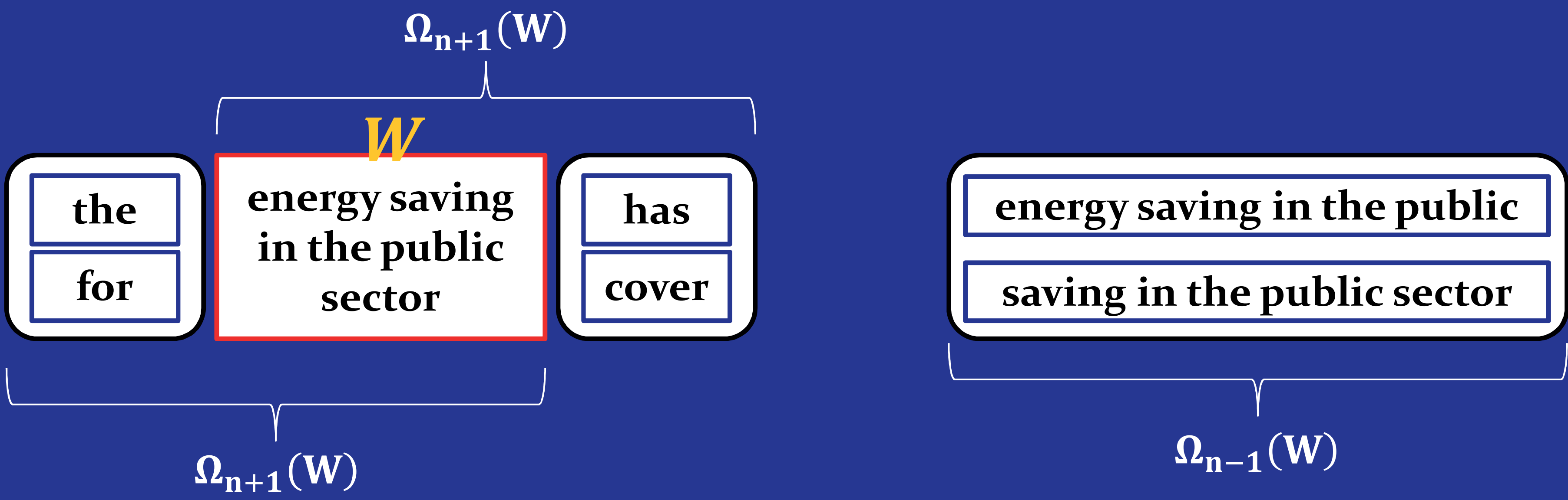
energy saving in the public sector

$$\underbrace{\textbf{saving in the public sector}}_{\Omega_{n-1}(W)}$$

LocalMaxs: Relevance Criterion

B) Higher *n*-grams: length(W) > 2

Comparing to the adjacent *n*-grams



$$\text{for all } \textit{glue}\ x \in \Omega_{n-1}(W), \text{ for all } \textit{glue}\ y \in \Omega_{n+1}(W)$$

$$SCP(W) > \frac{x+y}{2}$$

→ **W is Relevant !**

LocalMaxs

Logical Phases

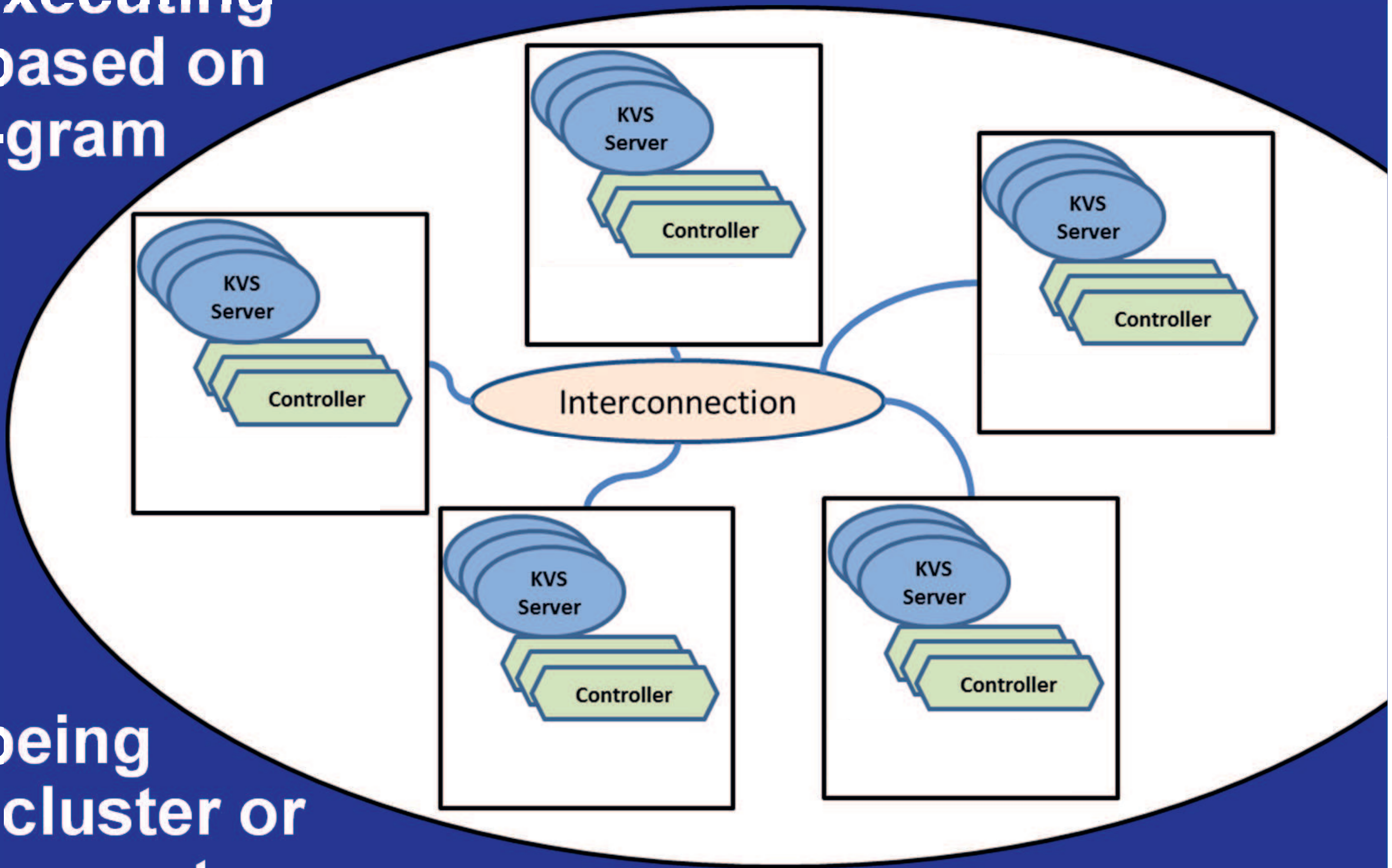
1) **Count all *n-gram* frequencies in the *corpus***
2) **Calculate all distinct *n-gram* glues (cohesion)**
3) **Find Relevant Expressions: for *n-gram* sizes from 2 to N, select the local stronger average glues**

Approaches

Sequential	Parallel & Distributed
Very time-consuming !!! →	Parallel: To reduce Time
Huge memory-demanding !!! →	Distributed: To fit in Memory

Distributed Architecture

- **Generic architecture capable of executing algorithms based on statistical *n*-gram models**

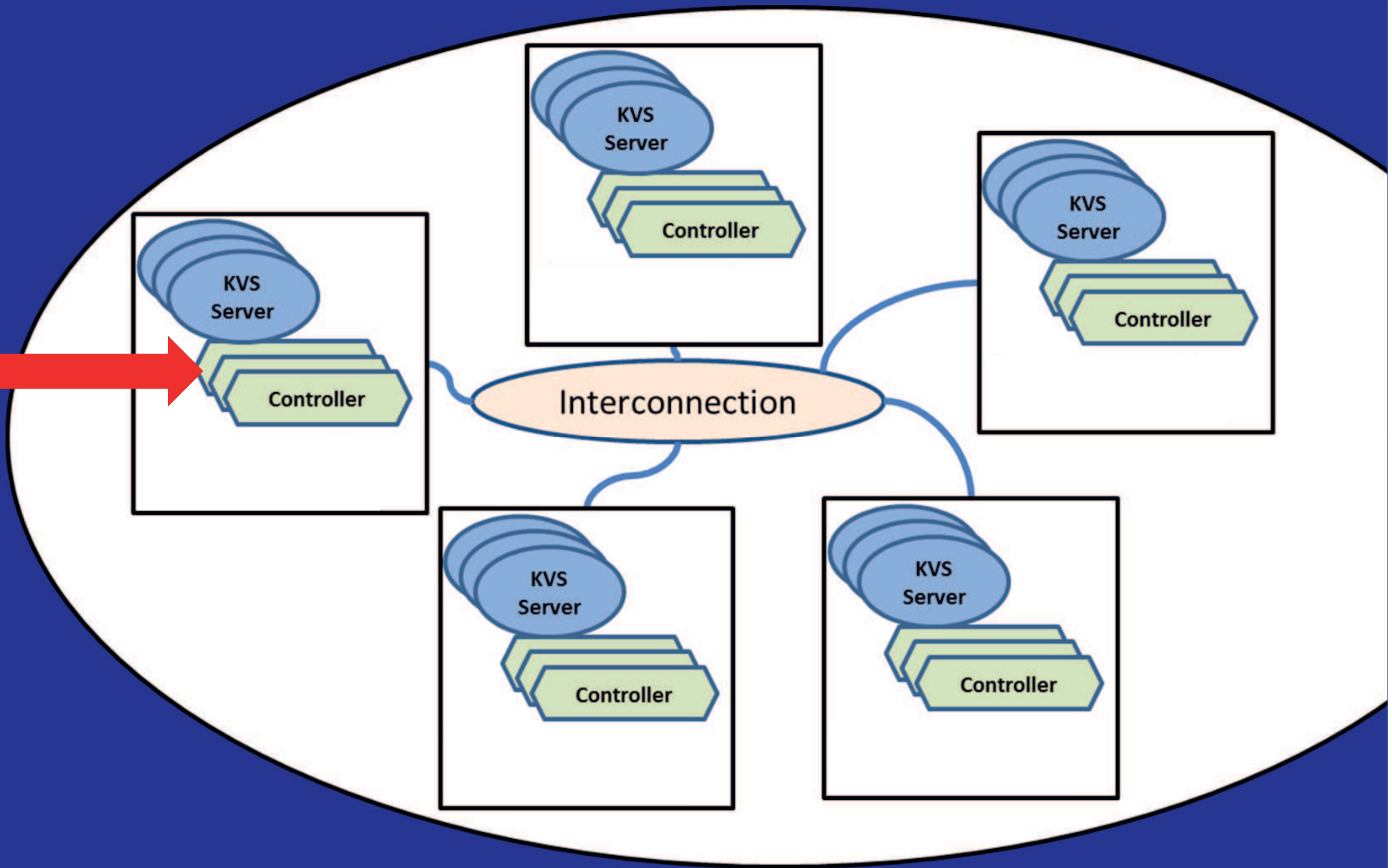


- **Capable of being executed in cluster or cloud environments**

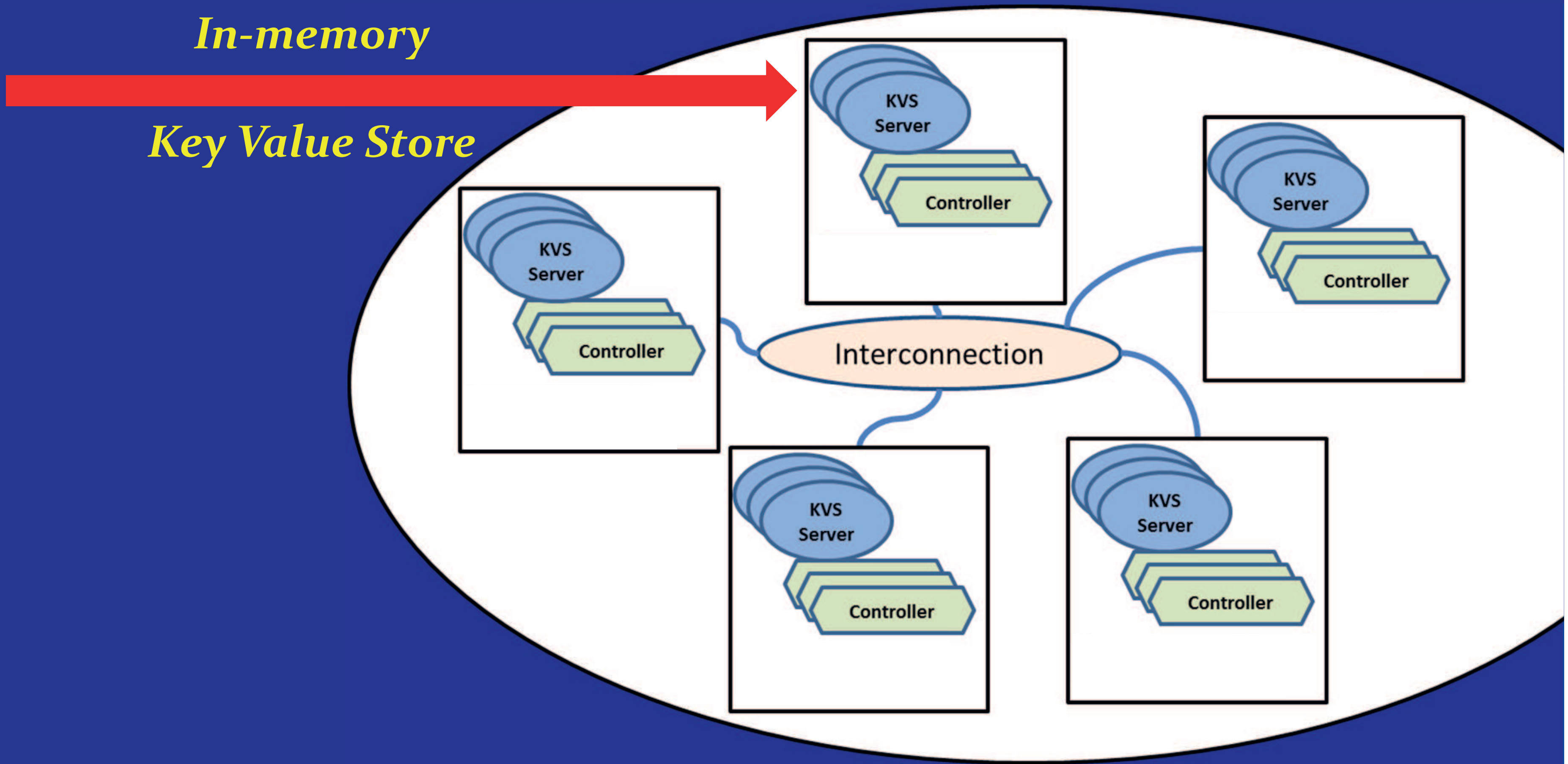
Distributed Architecture

LocalMaxs

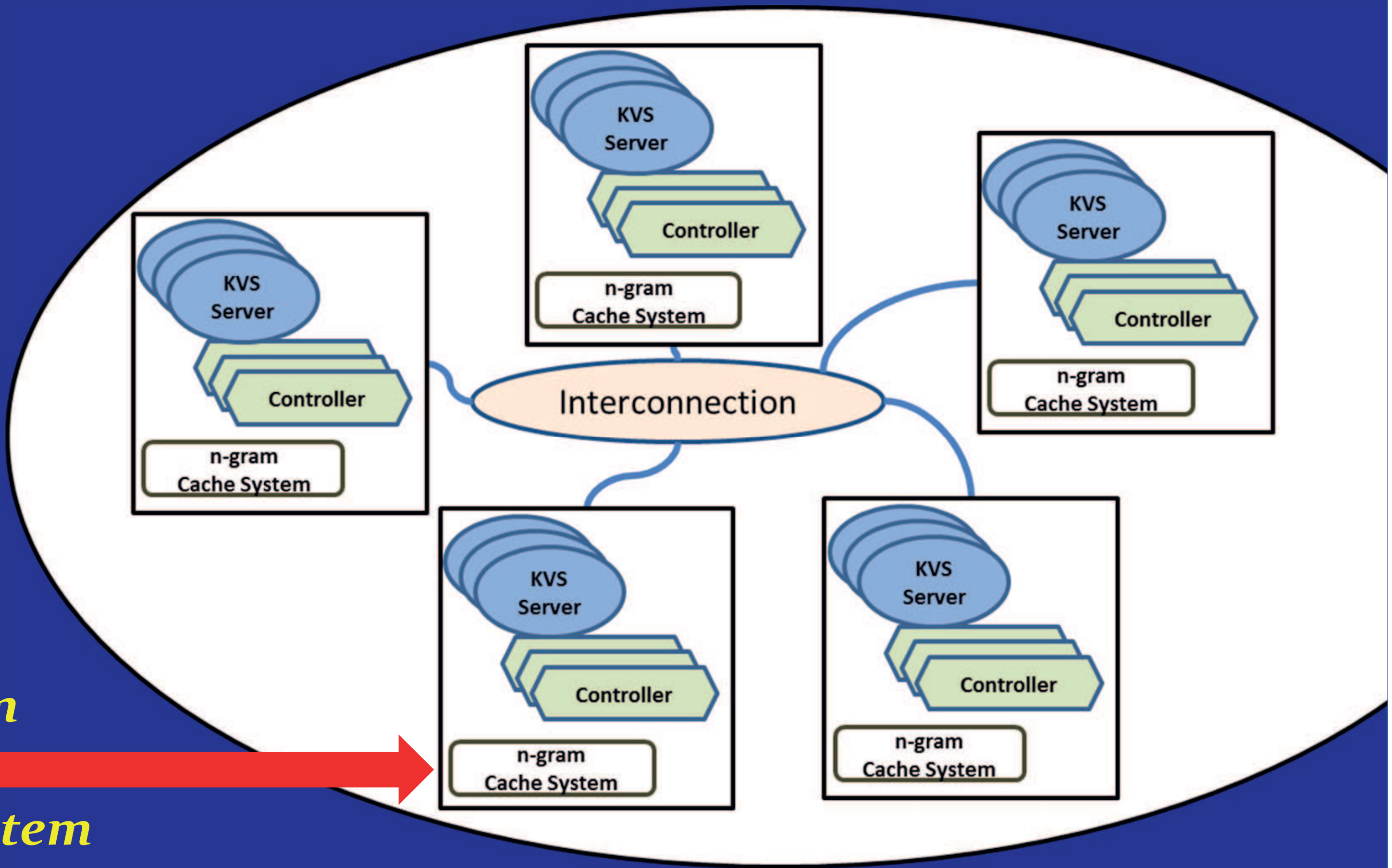
functions



Distributed Architecture



Distributed Architecture

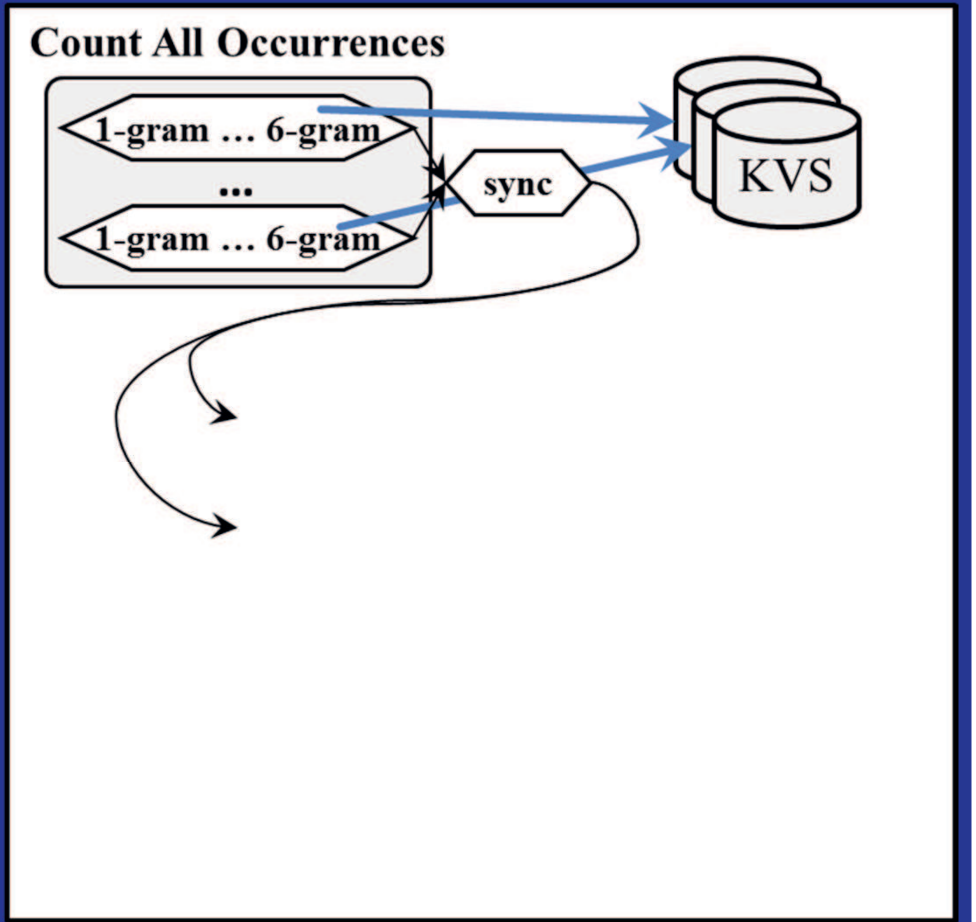


n-gram

Cache system

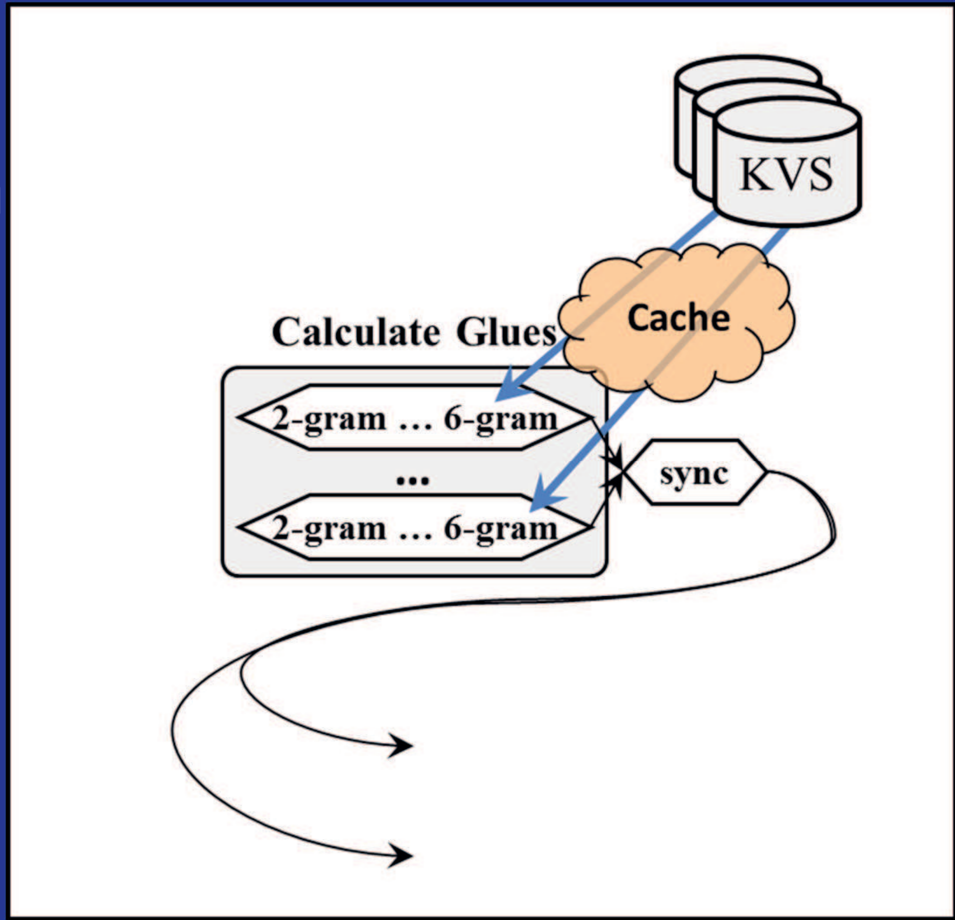
Distributed Architecture

- **Phase 1 counts the *n*-gram occurrences**
    - Distributed hash table with the *n*-gram data
- **Phase 2 calculate the cohesion**
- **Phase 3 identifies the *n*-grams that can be considered Relevant Expressions**



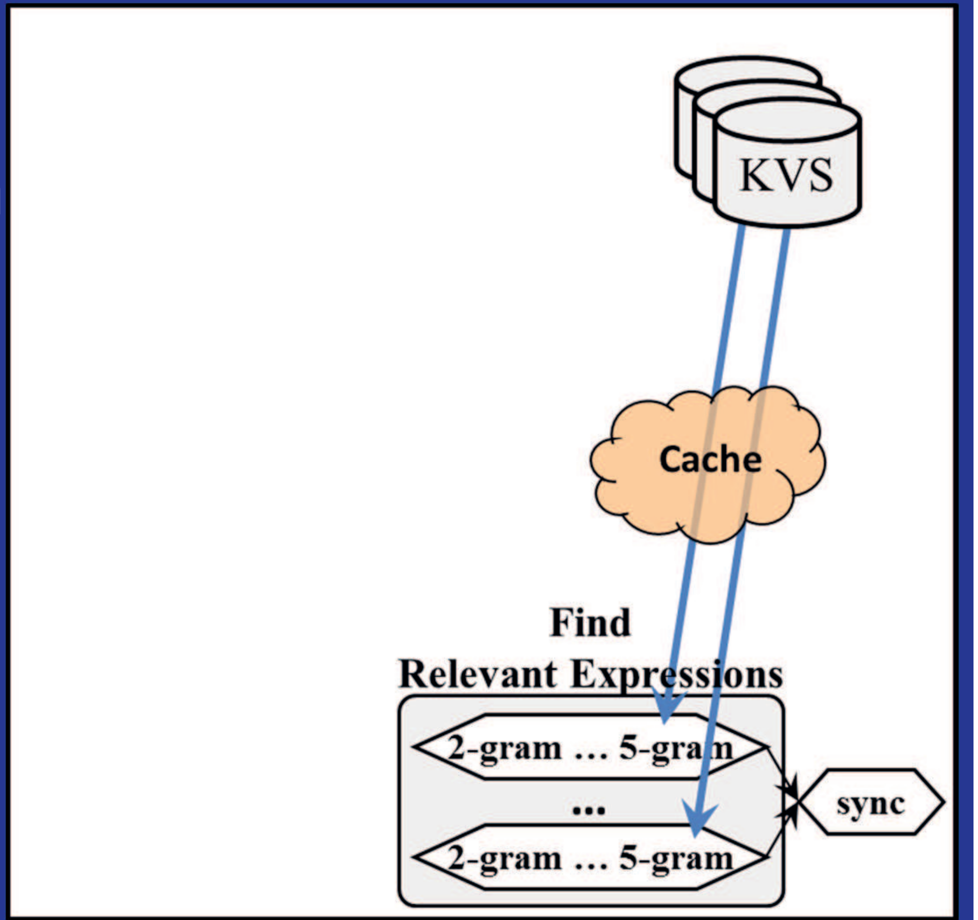
Distributed Architecture

- **Phase 1 counts the *n*-gram occurrences**
  - Distributed hash table with the *n*-gram data
- **Phase 2 calculate the cohesion**
- **Phase 3 identifies the *n*-grams that can be considered Relevant Expressions**



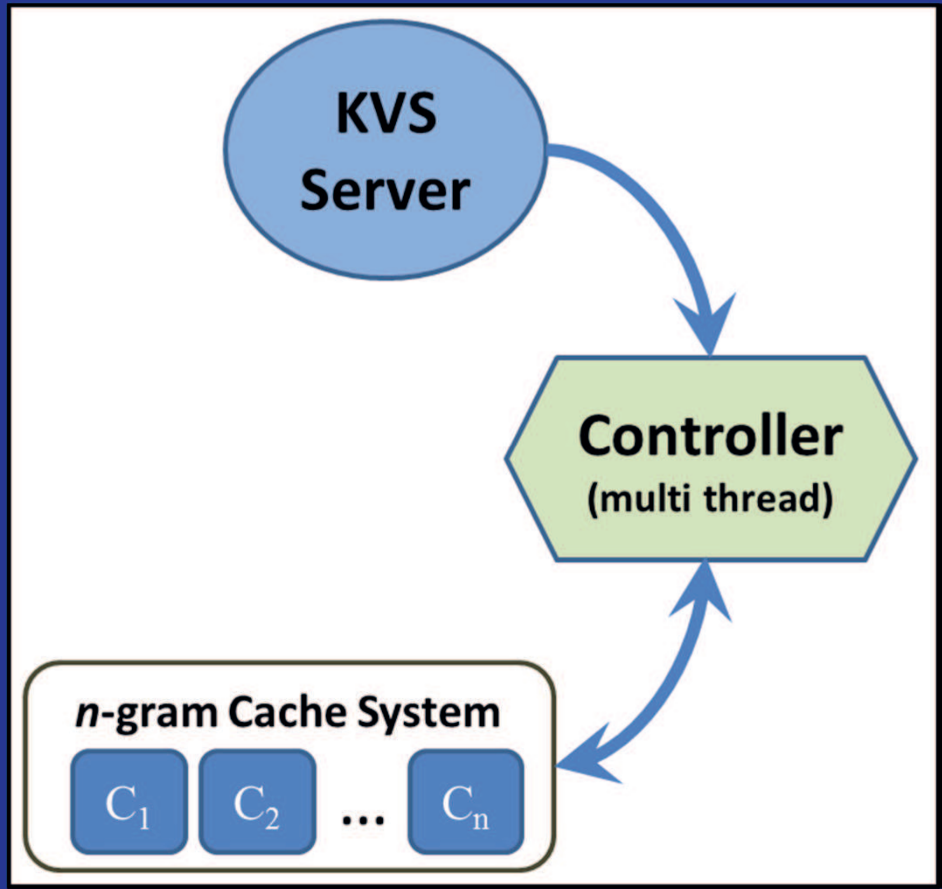
Distributed Architecture

- **Phase 1 counts the *n*-gram occurrences**
  - Distributed hash table with the *n*-gram data
- **Phase 2 calculate the cohesion**
- **Phase 3 identifies the *n*-grams that can be considered Relevant Expressions**



n-gram Cache System

- An *n*-gram cache system, to reduce the remote data communication
- Analytical model to understand cache miss ratio and miss penalty
- Cooperative warm-up strategy
- Finite size or *infinite*, depending on algorithm requirements or system resources

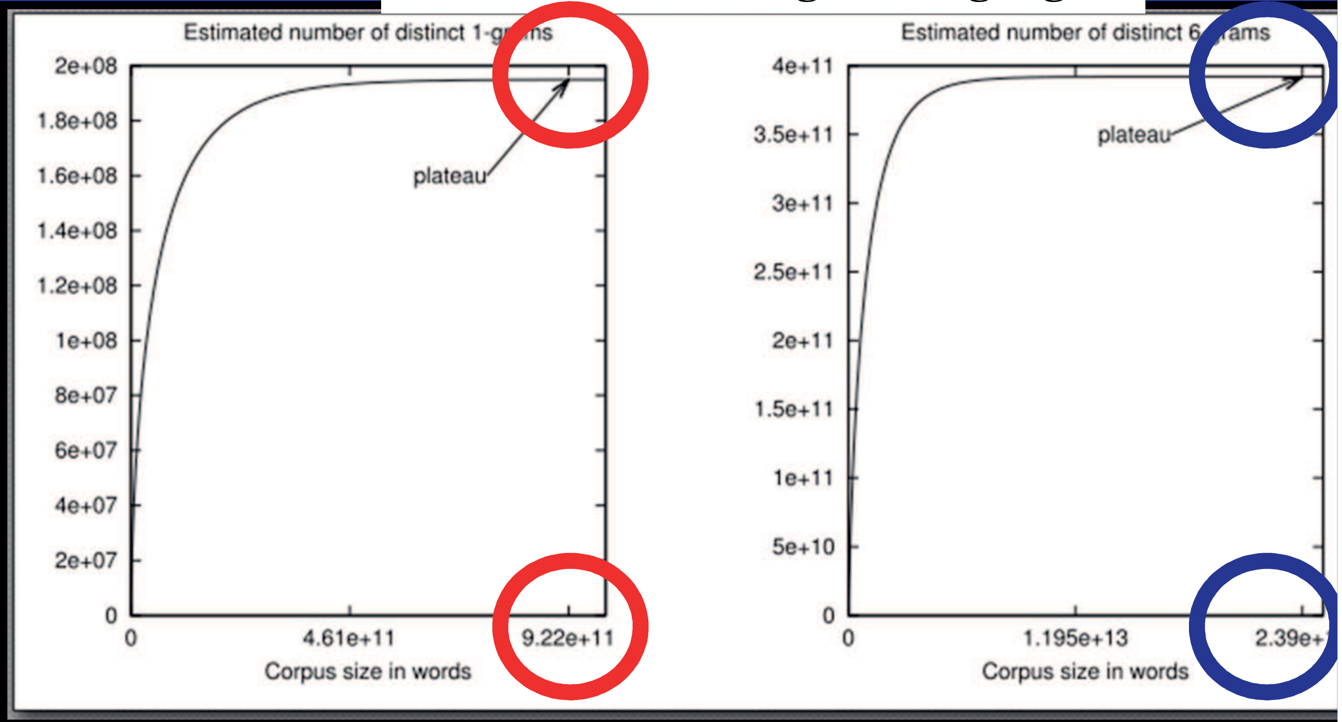


Enough to contain all distinct *n*-grams

n-grams Statistical Distribution

- ***n*-gram repetition depends on:**
  - Corpus size
  - Language
  - *n*-gram size
- **Empirical and Theoretical approaches**

***Plateaux* for the English language**



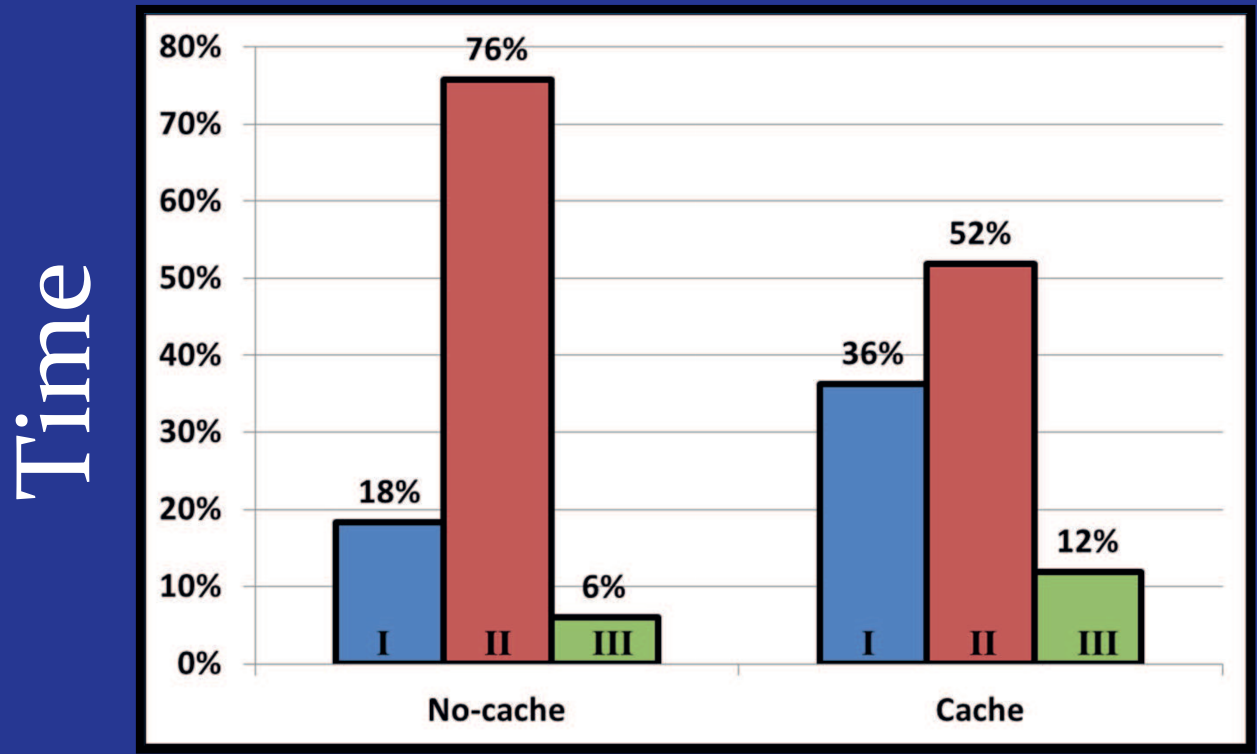
English *plateaux*

	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
D_i	1.95×10^8 (0.2 *Gw*)	7.08×10^8 (0.7 *Gw*)	3.54×10^9 (3.5 *Gw*)	9.80×10^9 (9.8 *Gw*)	5.06×10^{10} (50 *Gw*)	3.92×10^{11} (0.4 *Tw*)
$\lvert C\rvert$	9.22×10^{11} (0.9 *Tw*)	1.05×10^{12} (1 *Tw*)	1.29×10^{12} (1.2 *Tw*)	1.43×10^{12} (1.4 *Tw*)	6.18×10^{12} (6.2 *Tw*)	2.39×10^{13} (24 *Tw*)

Experimental Results

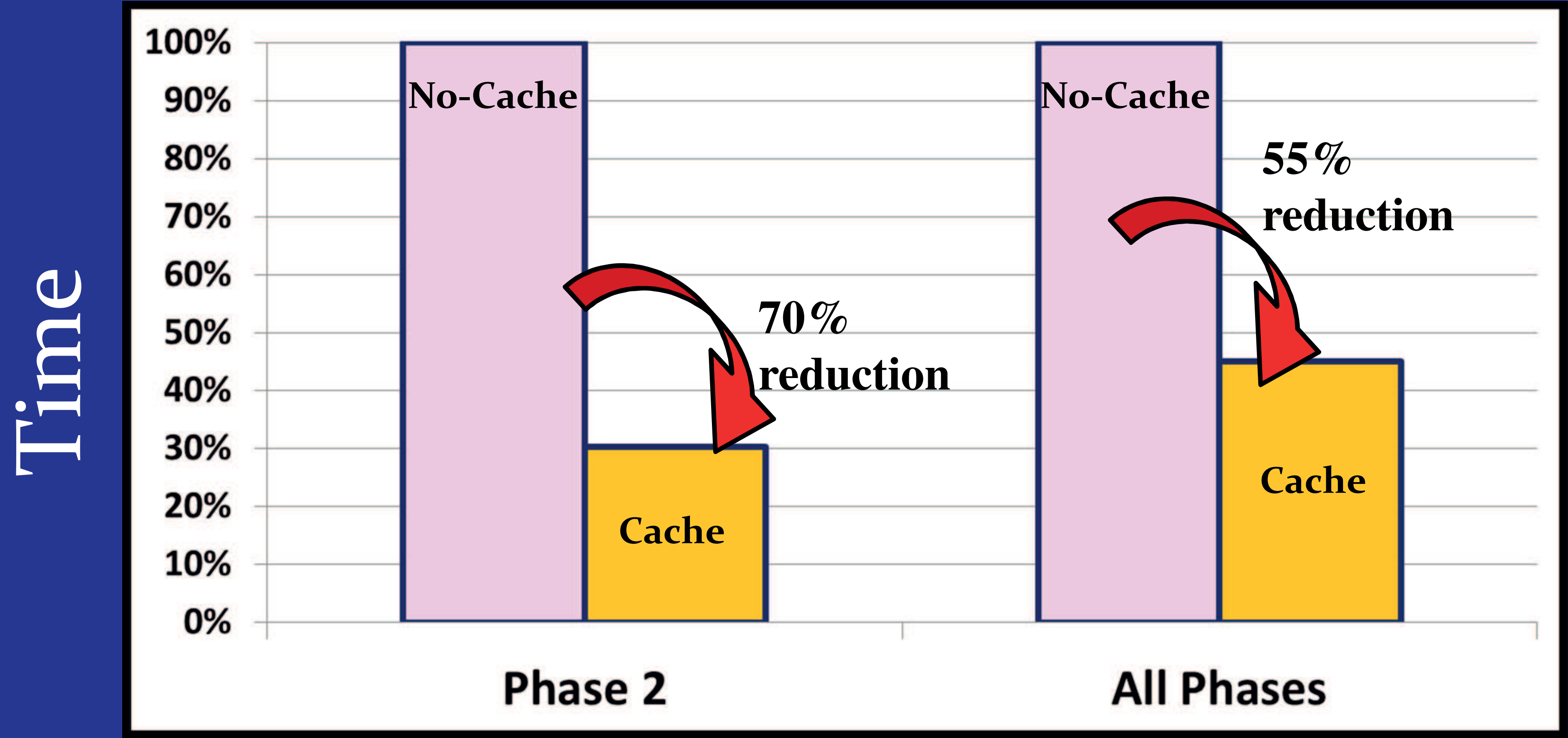
- **Multiple runs in public cloud (Lunacloud): virtual machines: 4 CPU@1.5 GHz and one local partition of 10 Gbyte**
- **Different number of machines (1, 9, 16, 24, 32, 40 and 48) with RAM ranging from 16 to 90 Gbyte, and different *corpus* sizes (25, 227, 466 and 682 million words)**
- **Evaluate:**
  - *n*-gram cache evaluation – Real execution results *vs* model estimates;
  - LocalMaxs phase 2 real execution time and cache behavior;
  - LocalMaxs total execution time *vs* phase 2 execution time

Experimental Results, 16 up to 48 virtual machines Corpus up to 682 Mw, *2*-grams & *3*-grams



- **Phase two execution time is dominated by the communication due to the *n*-gram misses in the observed range of *corpora* size and number of machines**

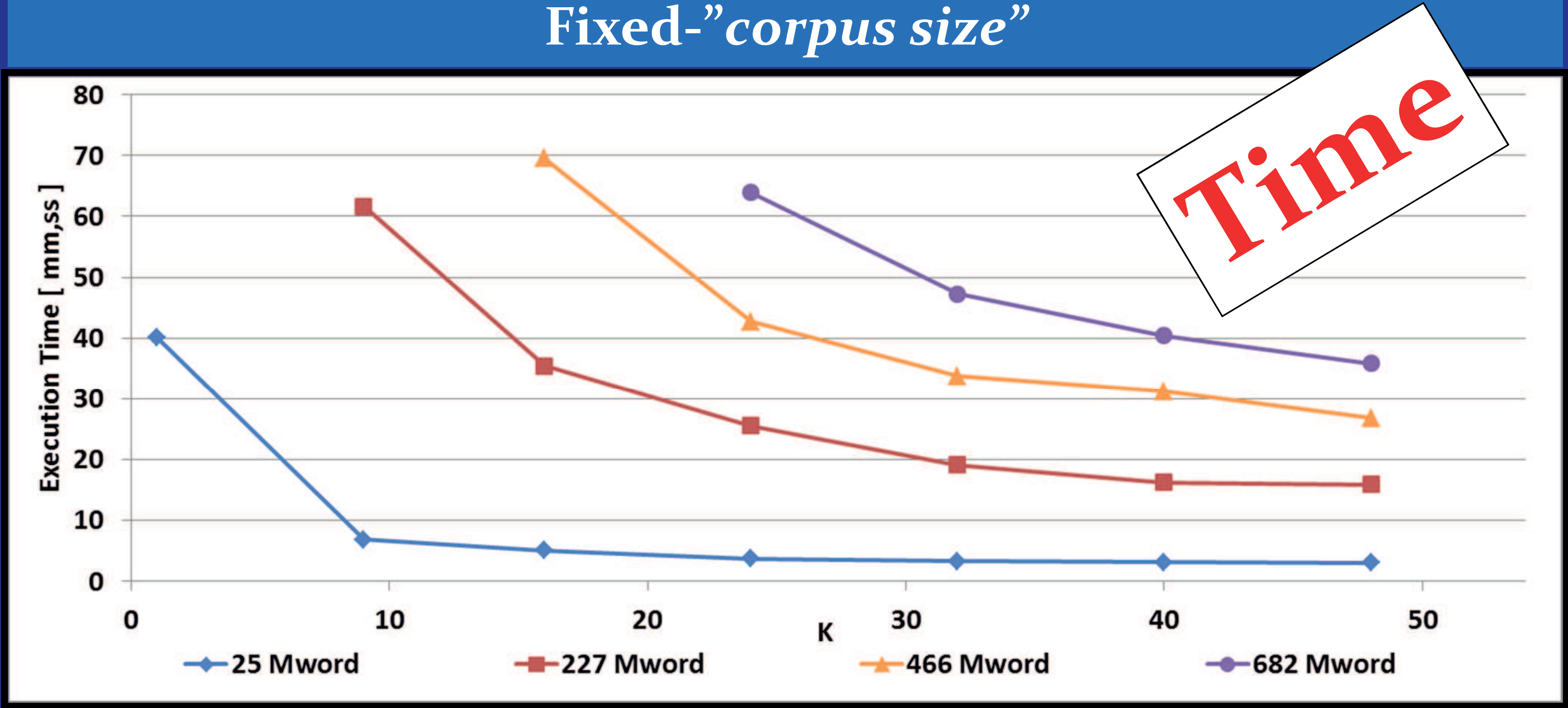
Experimental Results, 16 up to 48 virtual machines Corpus up to 682 Mw, *2*-grams & *3*-grams



- **Cache miss ratio ≈ 30%**
- **As *corpus* sizes increases the miss ratio decreases due to the repetition of *n*-grams**

Experimental Results
Extraction of relevant *2*-grams and *3*-grams

Fixed-"*corpus size*"

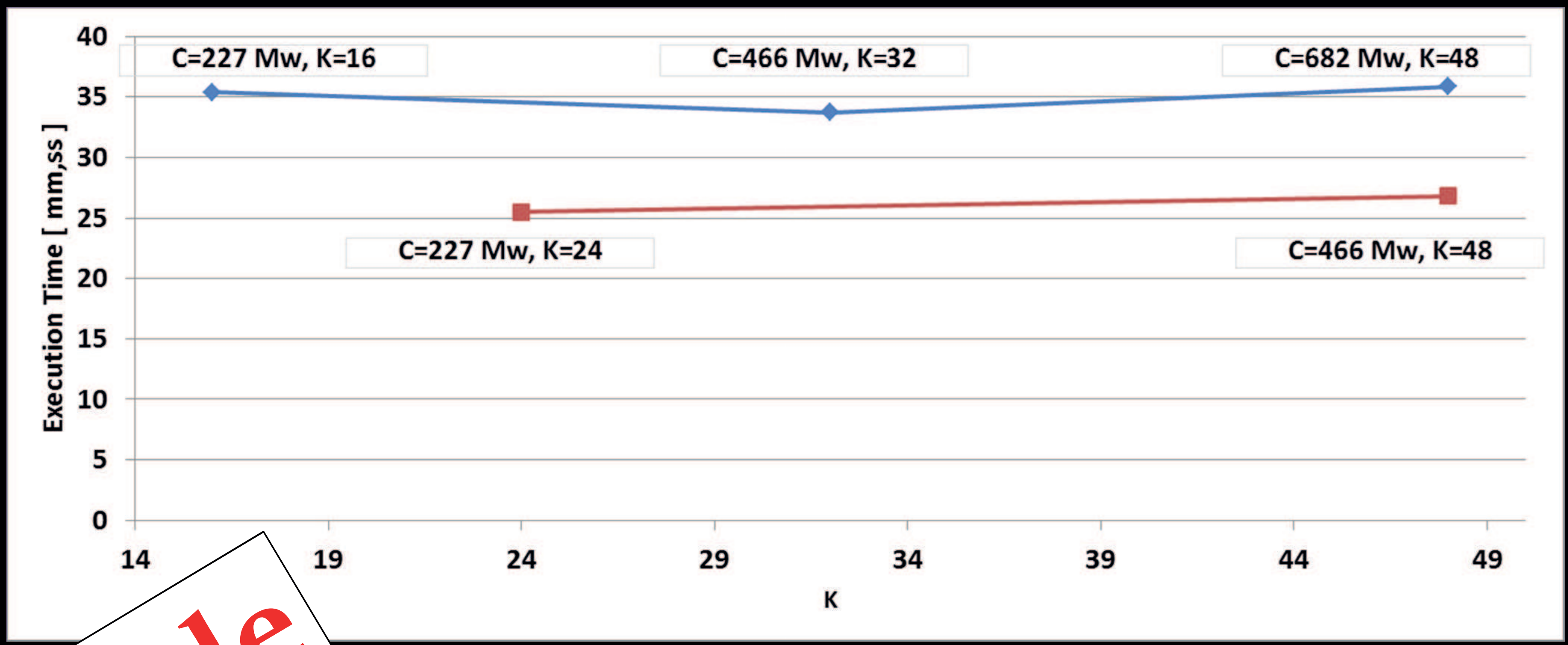


$$T \propto \frac{1}{K}$$

$$Sp_{K_1 \to K_2} = T_{par}(K_1)/T_{par}(K_2)$$

Experimental Results
Extraction of relevant *2*-grams and *3*-grams

Fixed-time Sizeup



Scale

$$Szp_{K_1 \rightarrow K_2} = N_{par}(T, K_2)/N_{par}(T, K_1)$$

Conclusions and Further Work

- **The approach is scalable to larger *corpora* sizes and higher size *n*-grams by simply increasing the number of machines**

- **An *n*-gram cache significantly reduced the remote data communication**

- **For each *corpus* size the number of distinct *n*-grams imposes a limit to the minimum remote communication overhead**

Thank you for your attention

Carlos Gonçalves (cgoncalves@deetc.isel.pt)