

عنوان ارائه:

مجموعه داده شغل‌های ابری گوگل برای زیرساخت‌های  
توزیع شده و رایانش ابری

**Google Cloud Jobs Dataset for Distributed and Cloud  
Computing Infrastructures**

توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۳۹۹/۱۲/۶

# مقدمه

## ■ اطلاعات مقاله

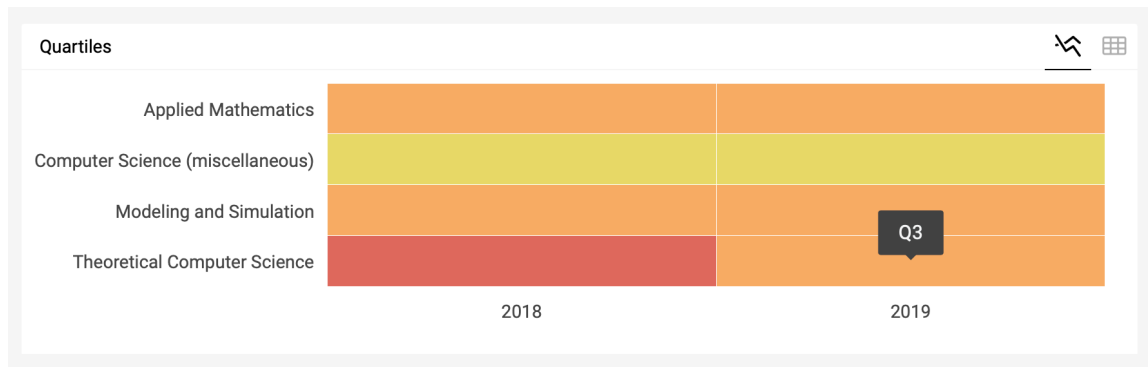
عنوان:

*GoCJ : Google Cloud Jobs Dataset for Distributed and Cloud Computing Infrastructures*

سال چاپ: 2018

تعداد ارجاع: 18

ناشر: *Multidisciplinary Digital Publishing Institute*



# مقدمه

## اطلاعات نویسندگان



### Altaf Hussain

KICSIT Campus, [Institute of Space Technology \(IST\)](#), Islamabad  
Verified email at mail.ist.edu.pk - [Homepage](#)

[Distributing Computing](#) [Cloud Computing](#) [Green Computing](#) [Data Mining](#) [Software Engineering](#)

FOLLOW

Citations	95	92
h-index	6	6
i10-index	4	3



### Muhammad Aleem

Professor at [National University of Computer and Emerging Sciences, Islamabad](#)  
Verified email at nu.edu.pk - [Homepage](#)

[Parallel Computing](#) [Cloud Scheduling](#) [Fog Computing](#) [Scheduling and Resource ...](#)

FOLLOW

Citations	309	287
h-index	10	9
i10-index	11	8

## فهرست مطالب

- مقدمه
- معرفی روش
- ارزیابی روش (داخلی)
- بررسی نقاط قوت و ضعف

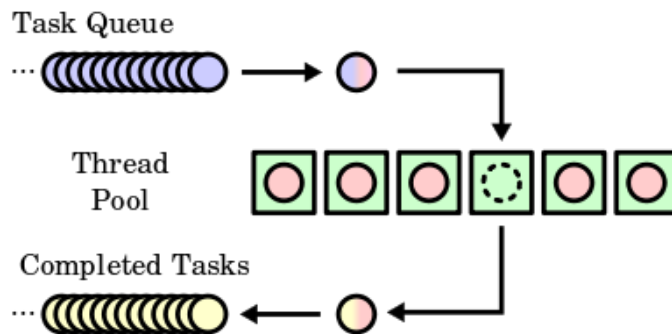
## مقدمه

### ■ الگوریتم‌های تخصیص منابع

- وظیفه تخصیص پویا و خودکار منابع به برنامه‌ها را بر عهده دارند
- در اقتصاد، برنامه‌ریزی استراتژیک، علوم کامپیوتر و غیره کاربرد دارد

### ■ الگوریتم‌های زمان‌بندی

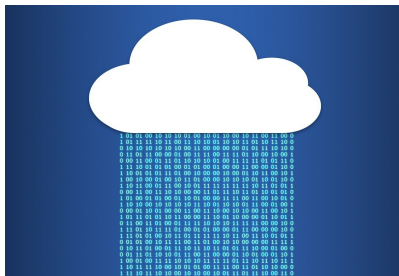
- وظیفه برنامه‌دهی به پردازنده‌ها برای تخصیص منابع را بر عهده دارند
- اهداف مختلفی مانند افزایش گذردهی، کاهش تاخیر (زمان پاسخ) را می‌توانند داشته باشند
- روش‌های عمومی متفاوتی دارد مانند *FIFO, SRT, RoundRobin* و غیره



## مقدمه

- چالش‌های ارزیابی کارایی الگوریتم‌های تخصیص‌دهی و زمان‌بندی
  - دسترسی به داده‌های واقعی ابری به دلیل سیاست‌های ارائه‌دهنده خدمات ابری و محرمانگی داده‌ها، سخت است
  - استفاده از بسترهای تست واقعی، آزمایش‌ها را به مقیاس آن محدود می‌کند
  - مجموعه داده‌های سنتز شده، هیچ‌گاه رفتار مجموعه داده‌های واقعی را از خود نشان نمی‌دهند

★ مناسب‌ترین کار این است که آزمایش‌ها در یک محیط شبیه‌سازی با بار رفتارهای متفاوت در محیط ابری انجام شود



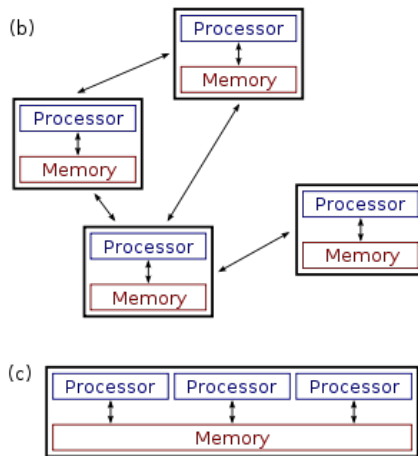
## مقدمه

■ محاسبات توزیع شده

■ از سیستم‌های توزیع شده برای حل مسائل استفاده می‌شود

■ مسئله به چندین وظیفه تقسیم می‌شود و بین سیستم‌های پردازشی پخش می‌شود

■ سیستم‌ها با یکدیگر به وسیله رد و بدل پیامی، در ارتباط هستند



## مقدمه

### ■ رایانش ابری

- ★ الگویی تازه برای عرضه، مصرف و تحویل خدمات رایانشی با به کارگیری شبکه است
- ★ مفهوم ابر به شبکه رایانه‌ای اشاره دارد و فضایی است که کاربر از پشت صحنه آن اطلاع دقیقی ندارد
- ★ مزایای رایانش ابری:

☑ دسترسی مقیاس پذیر پویا به مزایای تکنولوژی بدون داشتن دغدغه از استقرار، نگهداری و عملیات زیرساخت فیزیکی

☑ ارائه خدمات به صورت بستر یا زیرساخت برای استقرار، اجرا و شبیه‌سازی بلادرنگ





## مقدمه

### ■ برنامه‌های حریص محاسبات

\* تجسم داده‌های ترافیک شبکه‌های بزرگ

\* مکانیزم‌های کنترل یادگیری چند نخه برای شبکه‌های عصبی

\* آزمایش‌های عملکرد در مرتب‌سازی ادغامی

\* مرتب‌سازی‌های ادغامی بازگشتی در پردازش کلان‌داده‌ها

\* موازی‌سازی الگوریتم‌های مرتب‌سازی ادغامی تغییر یافته

★ نیاز به آرشیو مجموعه داده‌هایی که رفتار واقعی حجم کار ابر را نشان می‌دهند بسیار حس می‌شود زیرا این مجموعه داده‌ها برای ارزیابی عملکرد مکانیزم‌های زمان‌بندی ارائه شده سایر محققین، به کار می‌رود

## مقدمه

### ■ شبیه‌سازی مونت کارلو

★ طبقه‌ای از الگوریتم‌های محاسبه‌گر که برای محاسبه نتایج خود بر نمونه‌گیری‌های

تکرارشونده تصادفی اتکا می‌کنند

★ الگوی مشخص رویکردهای مونت کارلو:

☑️ تعریف محدوده‌ای از ورودی‌های ممکن

☑️ تولید ورودی‌های تصادفی از آن محدوده

☑️ انجام محاسبات دلخواه بر روی ورودی‌ها

☑️ ادغام نتایج هر کدام از اجراها در پاسخ نهایی

## معرفی روش

چند مجموعه داده در دسترس عموم ★

*Heterogeneous Computing Scheduling Problems (HCSP) Instances* ✓

*Task Execution Time Modeling (TETM)* ✓

*Google cluster traces* ✓

*Yahoo cluster traces* ✓

*Facebook Hadoop Workload* ✓

*OpenCloud Hadoop Workload* ✓

*Eucalyptus IaaS cloud Workload* ✓

*Grid Workload Archiver TuDelft (GWA – T) traces* ✓

## معرفی روش

ویژگی مجموعه داده جدید معرفی شده چیست؟

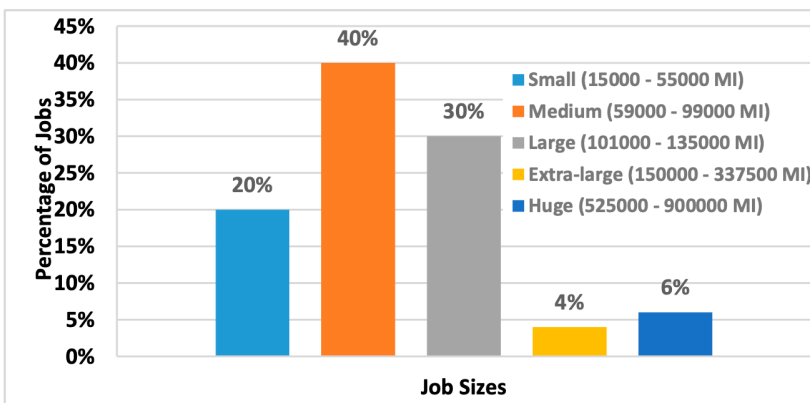
✓ مجموعه داده *GoCJ* بازتابی از رفتار بار واقعی را نشان می دهد که در ردیابی های خوشه گوگل و گزارش های نگاشت-کاهش خوشه ابررایانه *M45* دیده می شود. بنابراین برای محققانی که در زمینه های برنامه های مبتنی بر ابر و زمان بندی خوشه ها، مشغول به فعالیت هستند، از اهمیت و کاربرد بیشتری برخوردار است

✓ مجموعه داده *GoCJ* می تواند به عنوان جایگزینی برای بنچمارک گرفتن از بار کاری مکانیزم های زمان بندی و تخصیص حافظه با استفاده از مشاغل *HPC* در محاسبات ابری باشد

✓ داده ها به دو روش تولید می شوند: فایل های اکسل و ابزار جاوا

## معرفی روش

- مجموعه داده معرفی شده در مخزن داده *Mendeley* وجود داشته و شامل ۲۱ فایل می باشد. شماره نوشته شده روی هر فایل، نشان دهنده تعداد مشاغل آن فایل است
- هر فایل شامل چندین ردیف است که در آن، هر ردیف یک عدد در خود دارد که نشان دهنده حجم کار در مقیاس *MI* است
- زمان اتمام کارها از توزیع دم طولانی پیروی می کند؛ ۹۰ درصد مشاغل به صورت میانگین در ۱.۶ دقیقه پایان می پذیرد. ۶ درصد مشاغل زیر ۵ دقیقه و ۴ درصد مابقی در ۱۵ دقیقه به اتمام می رسد. بیشینه زمان اجرا ۱۵ دقیقه است. میانگین زمان اجرا نیز، ۵ دقیقه است



## معرفی روش

🌐 تجزیه و تحلیل منابع مورد بررسی قرار داده شده:

📍 ردیاب‌های خوشه گوگل

✅ ردیاب‌های ۲۹ روز مورد بررسی و آزمایش قرار گرفتند

✅ نتایج نشان داده است که اکثر شغل‌ها در کمتر از ۱۵ دقیقه پایان می‌پذیرند. درصد بسیار

کمی حتی بیش از ۳۰۰ دقیقه هم طول کشیدند. میانه طول یک شغل نیز تقریباً ۳ دقیقه تخمین زده شده است

✅ بنابراین دو سوم مشاغل در کمتر از ۵ دقیقه و تقریباً، ۲۰ درصد آن‌ها در کمتر از یک دقیقه

اجرا می‌شوند

★ بنابراین در ردیاب‌های خوشه گوگل، اکثر مشاغل طول اجرای کوتاهی دارند و به صورت

کلی می‌توان گفت که مشاغل کوچک برای آزمایش بر روی خوشه گوگل انتخاب و استفاده

می‌شوند

# معرفی روش

🌐 تجزیه و تحلیل منابع مورد بررسی قرار داده شده (ادامه):

📌 گزارش‌های نگاشت-کاهش ابرایانه  $M45$

✅ گزارش‌های ۱۰ ماه مورد بررسی و آزمایش قرار گرفتند (منتشر شده به وسیله یاهو)

✅ نتایج نشان داده است که اکثر شغل‌ها (۹۵ درصد)، در حدود ۲۰ دقیقه اجرا می‌شوند و تقریباً

۴-۵ درصد مابقی، تا ۳۰ دقیقه هم به طول می‌انجامد

★ براساس موارد بررسی شده، مجموعه‌داده تحقق‌گرای  $GOCJ$  به وسیله شبیه‌سازی مونت کارلو تولید می‌شود. به جای استفاده از روش تولید اعداد تصادفی ( $RNG$ )، مجموعه‌داده اصلی توسط روش فوق، به صورت مکرر نمونه‌گیری می‌شود

✅ هر اندازه شغل در نمونه‌گیری پرتکرار، احتمال یکسان با بقیه دارد

✅ فرض شده است که میانگین قدرت ماشین در محیط محاسباتی توزیع شده، هزار  $MI$  بر ثانیه

است

# معرفی روش

اندازه مشاغل در مجموعه داده اصلی *GoCJ*

Table 3. Sizes of jobs in original dataset for GoCJ (in MIs).

Small	Medium	Large	Extra-Large	Huge
	59,000, 61,000,			
	63,000, 65,000,	101,000, 103,000,		
	67,000, 71,000,	105,000, 107,000,		
15,000, 27,500,	73,000, 75,000,	109,000, 111,000,		
40,000, 45,000,	77,000, 79,000,	113,000, 115,000,		525,000,
47,000, 49,000,	81,000, 83,000,	117,000, 119,000,	150,000, 525,000	712,500,
51,000, 53,000,	85,000, 87,000,	121,000, 123,000,		900,000
55,000	89,000, 91,000,	125,000, 127,000,		
	93,000, 95,000,	129,000, 135,000		
	97,000, 99,000			


\* نحوه تبدیل واحدهای زمان مورد انتظار برای اتمام (*ETC*) و میلیون دستورالعمل (*MI*)

$$* ETC_{Second} = \frac{Job_{MI}}{Machine_{MIPS}}$$

$$* Job_{MI} = Machine_{MIPS} \times ETC_{Second}$$



# معرفی روش

روش‌های بازتولید مجموعه داده 

	A	B	C	D	E	F	G
1	0.02	0	15000		0.82477	125000	
2	0.02	0.02	27500		0.0067	15000	
3	0.02	0.04	40000		0.23684	63000	
4	0.02	0.06	45000		0.92653	337500	
5	0.02	0.08	47000		0.89473	135000	
6	0.02	0.1	49000		0.72494	115000	
7	0.02	0.12	51000		0.91522	150000	
8	0.02	0.14	53000		0.06835	45000	
9	0.02	0.16	55000		0.02644	27500	
10	0.02	0.18	59000		0.15855	53000	
11	0.02	0.2	61000		0.41961	83000	
12	0.02	0.22	63000		0.0429	40000	
13	0.02	0.24	65000		0.32682	75000	
14	0.02	0.26	67000		0.77935	119000	
15	0.02	0.28	71000		0.08815	47000	
16	0.02	0.3	73000		0.40771	83000	
17	0.02	0.32	75000		0.66957	109000	
18	0.02	0.34	77000		0.9136	150000	
19	0.02	0.36	79000		0.35568	77000	
20	0.02	0.38	81000		0.51804	93000	
21	0.02	0.4	83000		0.29206	71000	
22	0.02	0.42	85000		0.90112	150000	
23	0.02	0.44	87000		0.20031	61000	
24	0.02	0.46	89000		0.82516	125000	
25	0.02	0.48	91000		0.41658	83000	
26	0.02	0.5	93000		0.0184	15000	
27	0.02	0.52	95000		0.98706	900000	
28	0.02	0.54	97000		0.51058	93000	
29	0.02	0.56	99000		0.80718	123000	
30	0.02	0.58	101000		0.19399	59000	
31	0.02	0.6	103000		0.84291	127000	
32	0.02	0.62	105000		0.04853	40000	
33	0.02	0.64	107000		0.65054	107000	
34	0.02	0.66	109000		0.66589	109000	
35	0.02	0.68	111000		0.00755	15000	
36	0.02	0.7	113000		0.49215	91000	
37	0.02	0.72	115000		0.99745	900000	

☆ ستون  $A$ ، احتمال رخداد هر شغل را نشان

می‌دهد

☆ ستون  $B$ ، احتمال تجمعی را نشان می‌دهد

☆ ستون  $C$ ، اندازه مشاغل جدول قبلی را نشان

می‌دهد (حتماً ۵۰ سطر دارد)


☆ ستون  $E$ ، یک عدد تصادفی یکنواخت از صفر تا

یک است

☆ ستون  $F$ ، نزدیکترین عدد به عدد ستون  $E$  در

احتمال‌های تجمعی است

# معرفی روش

روش‌های بازتولید مجموعه داده (ادامه) 

---

## Algorithm 1: GoCJ Generator

---

**Input:** num — desired number of jobs in dataset,  
Original\_DataSet — file of original dataset sample

**Output:** jList — list of job sizes in the desired dataset

```
1 cPer = 0
2 jobSize = 0
3 jList = Null
4 dataTable < cPer, jobSize ≥ Null
5 fileReader = readFile(Original_DataSet)
6 bufferReader = read(fileReader)
7 while bufferReader is Not Empty do
8     jobSize = long.parseLong(bufferReader.readLine())
9     dataTable.add(cPer, jobSize)
10    cPer = cPer + 2
11 a=1
12 while num ≥ a do
13     rand = Random.nextInt(100)
14     jList.add((rand Mod 2)?dataTable.get(rand): getJobSize(rand))
15     a++
16 return jList
```

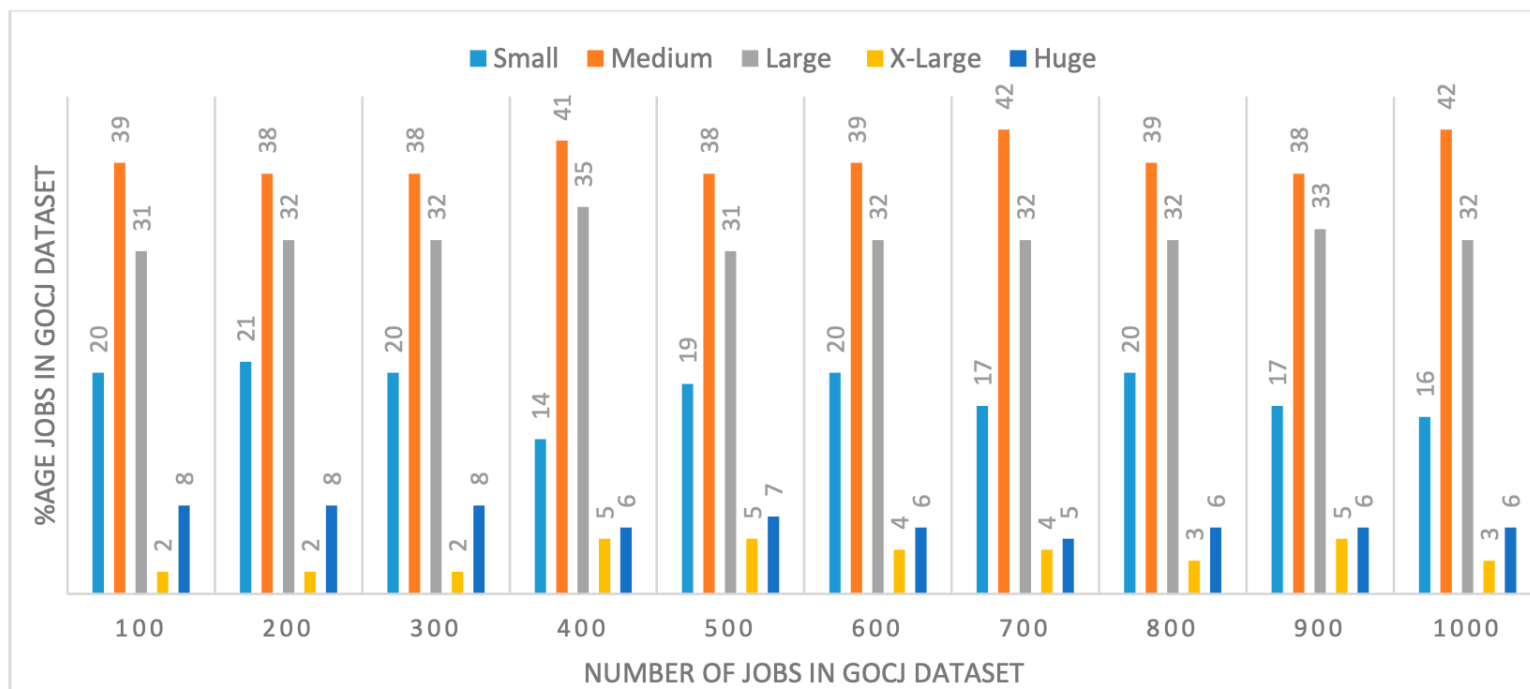
---

پیچیدگی الگوریتم، از  
مرتبه  $O(n)$  می‌باشد

# ارزیابی داخلی

بررسی انطباق با مجموعه داده اصلی

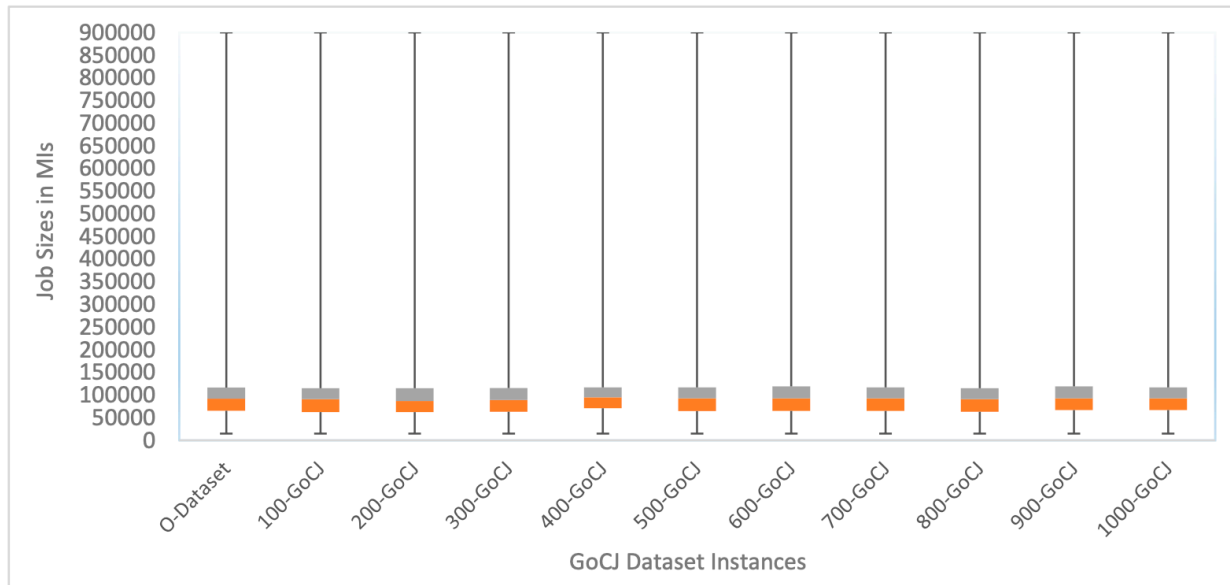
★ آزمون کوواریانس (مقدار برابر ۲.۴۹)



# ارزیابی داخلی

بررسی انطباق با مجموعه داده اصلی (ادامه)

☆ مصورسازی جعبه‌ای براساس ۵ عدد (کمینه، چارک اول، میانه، چارک سوم و بیشینه)



## بررسی نقاط قوت و ضعف

### ☆ نقاط قوت

- ☑ یک مجموعه داده واقع‌گرا با اندازه‌های گوناگون است
- ☑ در زمان‌بندی ابر، سیاست‌های تخصیص حافظه و آنالیزهای کارایی مبتنی بر بنچمارک می‌تواند مورد استفاده قرار بگیرد

### ☆ نقاط ضعف

- ☑ فقط براساس یک زیرساخت (ردیاب‌های خوشه گوگل) است
- ☑ شامل شغل‌های ضرب‌الاجل محور و توافق‌نامه سطح خدمات محور نمی‌باشد

## با تشکر از توجه شما