Gmail **Chaser Remmen <gcp@olderguyai.com>**

# Re: Older Guy AI & GCP Team

7 messages

**Chaser Remmen** <chaser@olderguyai.com> Thu, Oct 24, 2024 at 11:12 AM
To: Brandon Velasquez <velasquezb@google.com>
Cc: Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>, gcp@olderguyai.com

NameIDBilling accountBilling account IDActions
a-d-a-p-t
a-d-a-p-t
My Billing Account
01E6A9-93B716-08CC16

This is the actual name and ID:

"My Billing Account" << Actual Name
01E6A9-93B716-08CC16Active$0Directolderguyai.com 3

Project info

Project name
a-d-a-p-t
Project number
231017561254
Project ID
a-d-a-p-t

Thanks Guys!!!

Chaser

On Thu, Oct 24, 2024 at 10:56 AM Brandon Velasquez <velasquezb@google.com> wrote:
>
> Hello Chaser,
>
> Can you please provide the Billing ID, Project ID, and Project Number for the quota increase?
>
> Thanks,
> Brandon
>
> On Wed, Oct 23, 2024 at 4:49 PM Chaser Remmen <chaser@olderguyai.com> wrote:
>>
>> Send it over a little bit ago confirm you got it please thanks chaser
>>
>> On Wed, Oct 23, 2024 at 1:55 PM Joe Herbert <josephherbert@google.com> wrote:
>>>
>>> Chaser - Are you able to send over the specs you mentioned ? And connect us
>>>
>>>
>>> Joe Herbert
>>>
>>> Cloud Account Manager
>>>
>>> Mobile 650-664-7392
>>>
>>>
>>>
>>>
>>>
>>> On Wed, Oct 23, 2024 at 1:37 PM Joe Herbert <josephherbert@google.com> wrote:

>>>>
>>>> Starting new thread here
>>>>
>>>> Joe Herbert
>>>>
>>>> Cloud Account Manager
>>>>
>>>> Mobile 650-664-7392
>>>>
>>>>
>>>>
>
>
> --
>
> Brandon Velasquez
>
> Customer Engineer
>
> 650-609-7175
>
> 901 E 5th St, Austin, TX 78702
>
>

---

**Brandon Velasquez** <velasquezb@google.com> Fri, Oct 25, 2024 at 8:53 AM
To: Chaser Remmen <chaser@olderguyai.com>
Cc: Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>, gcp@olderguyai.com

Hello Chaser,

I have the quota increase request in progress for the GPU, storage, and network requirements. Do you have additional details for the C4A highmem? The team is requesting the region, total number of cores, shape, and LSSD requirements.

Best,
Brandon
[Quoted text hidden]
[Quoted text hidden]

---

**Chaser Remmen** <gcp@olderguyai.com> Fri, Oct 25, 2024 at 1:42 PM
To: Brandon Velasquez <velasquezb@google.com>
Cc: Chaser Remmen <chaser@olderguyai.com>, Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>

Hi Brando,

Here are the numbers:

In addition to the below, I need the following vm shapes and GPUs, if possible,fo for ongoing operations:

family - 50 ea (as we will mix and match)
T2a
N2
N2d
N4
C2
C3
C4
C4a (already in request)
H3
M1
M2
M3 (already in request)

A2
A3 (already in request)
X4
Z3

GPU (80 Ea)
H100 - 80MB
H100 - 80MB
A100 - 80MB
A100 - 40MB

Info Requested
Regions: us-central1, us-east4

C4a
# Cores: 9280
# C4a VMs - 200
Storage;
LSSD = 2PB
Hyperdisk Extreme = 2PB
Hyperdisk Balanced = 2PB

## C4A Highmem VM Shapes:

| VM Shape | CPU per VM | # Vms | CPU per Shape | Total |
|---|---|---|---|---|
| c4a-highmem-16 | 16 | 20 | 320 | |
| c4a-highmem-32 | 32 | 20 | 640 | |
| c4a-highmem-48* | 48 | 20 | 960 | |
| c4a-highmem-64* | 64 | 20 | 1280 | |
| c4a-highmem-72 | 72 | 20 | 1440 | **4640** |

| VM Shape | CPU per VM | # Vms | CPU per Shape | Total |
|---|---|---|---|---|

## C4A Highmem with Local SSD VM Shapes:

| | CPU per VM | # Vms | CPU Totals | Grand Total | |
|---|---|---|---|---|---|
| c4a-highmem-16-lssd | 16 | 20 | 320 | | |
| c4a-highmem-32-lssd | 32 | 20 | 640 | | |
| c4a-highmem-48-lssd* | 48 | 20 | 960 | | |
| c4a-highmem-64-lssd* | 64 | 20 | 1280 | | **Total CPUs Needed** |
| c4a-highmem-72-lssd | 72 | 20 | 1440 | **4640** | **9280** |

[Quoted text hidden]

---

**Chaser Remmen** <gcp@olderguyai.com> Fri, Oct 25, 2024 at 2:17 PM
To: Brandon Velasquez <velasquezb@google.com>
Cc: Chaser Remmen <chaser@olderguyai.com>, Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>

Brandon, (sorry for the typo last email)

Can you please add

### Extended Quota Increase Request:

1. **Cloud Storage**:
    - **Standard Storage**: Up to **2PB** for hot data requiring high throughput and frequent access.
    - **Nearline Storage**: Up to **1PB** for warm data with lower access frequency.
    - **Coldline Storage**: Up to **2PB** for low-cost storage of infrequently accessed datasets.
    - **Archive Storage**: Up to **1PB** for deep archival storage of long-term data with minimal access.
2. **Persistent Disk SSD**:
    - **SSD Storage**: Up to **2PB** for high-performance storage with fast read/write speeds to support critical operations.

[Quoted text hidden]

---

**Brandon Velasquez** <velasquezb@google.com> Mon, Oct 28, 2024 at 11:42 AM
To: Chaser Remmen <gcp@olderguyai.com>
Cc: Chaser Remmen <chaser@olderguyai.com>, Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>

Hello Chaser,

I am still working on your request internally. The capacity team wanted some additional information for the A3 GPUs:

1. Are you flexible on the region for the A3 GPUs?
2. Can your workload be split across multiple zones?
3. Are you planning on leveraging any CUDs (1yr or 3yr)?
4. Is your team open to using L4s for your workload?

Best,
Brandon

[Quoted text hidden]

---

**Chaser Remmen** <gcp@olderguyai.com> Mon, Oct 28, 2024 at 11:57 AM
To: Brandon Velasquez <velasquezb@google.com>
Cc: Chaser Remmen <chaser@olderguyai.com>, Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>

Brandon,

Answers below:

1. Are you flexible on the region for the A3 GPUs? --- Yes
2. Can your workload be split across multiple zones? --- Yes
3. Are you planning on leveraging any CUDs (1yr or 3yr)? --- yes, if this works out
4. Is your team open to using L4s for your workload? - No, too slow for our needs, would be open to the TPUv5p's

on the GPU/TPU - speed is critical for our use case, not only for training/fine tuning, but also for our Agentic workflows.

If you could get us a partial, while putting this together, it would help us get some things moving...we need to DL over 1000 LLM Models and 500+ Large Data Sets...if we could get a jump on that it would help us a lot. - LMK

Let me Know if you have any questions...

Chaser
[Quoted text hidden]

---

**Brandon Velasquez** <velasquezb@google.com> Thu, Oct 31, 2024 at 9:16 AM

To: Chaser Remmen <gcp@olderguyai.com>
Cc: Chaser Remmen <chaser@olderguyai.com>, Joe Herbert <josephherbert@google.com>, Mo Farhat <mofarhat@google.com>, nbaum@google.com, Ning Yang <ningyang@google.com>

Hello Chaser,

The capacity team is only able to support a partial grant of 1,500 cores of c4a-highmem at this time. Would your team be able to use these in combination with c4a-standard to run your workload?

Best,
Brandon
[Quoted text hidden]