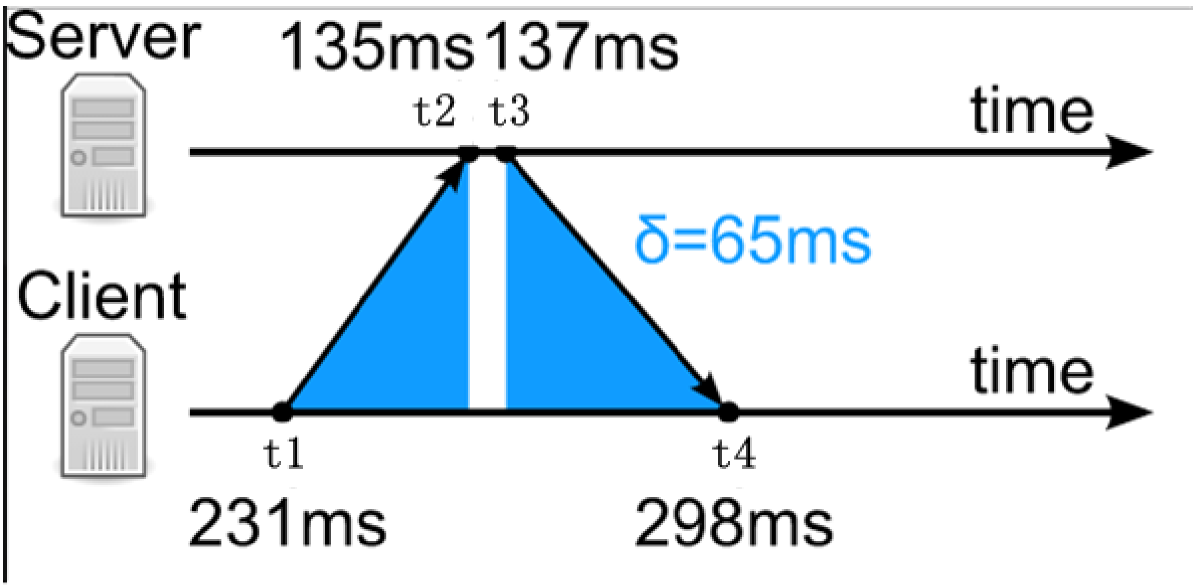
TCS_RDMA 时延统计量

1 前情提要

1.1 NTP



- $\sigma = (t_4 - t_1) - (t_3 - t_2) = \{\text{blue part}\}$
- $\delta = t_2 - t_1 - \frac{\sigma}{2}$

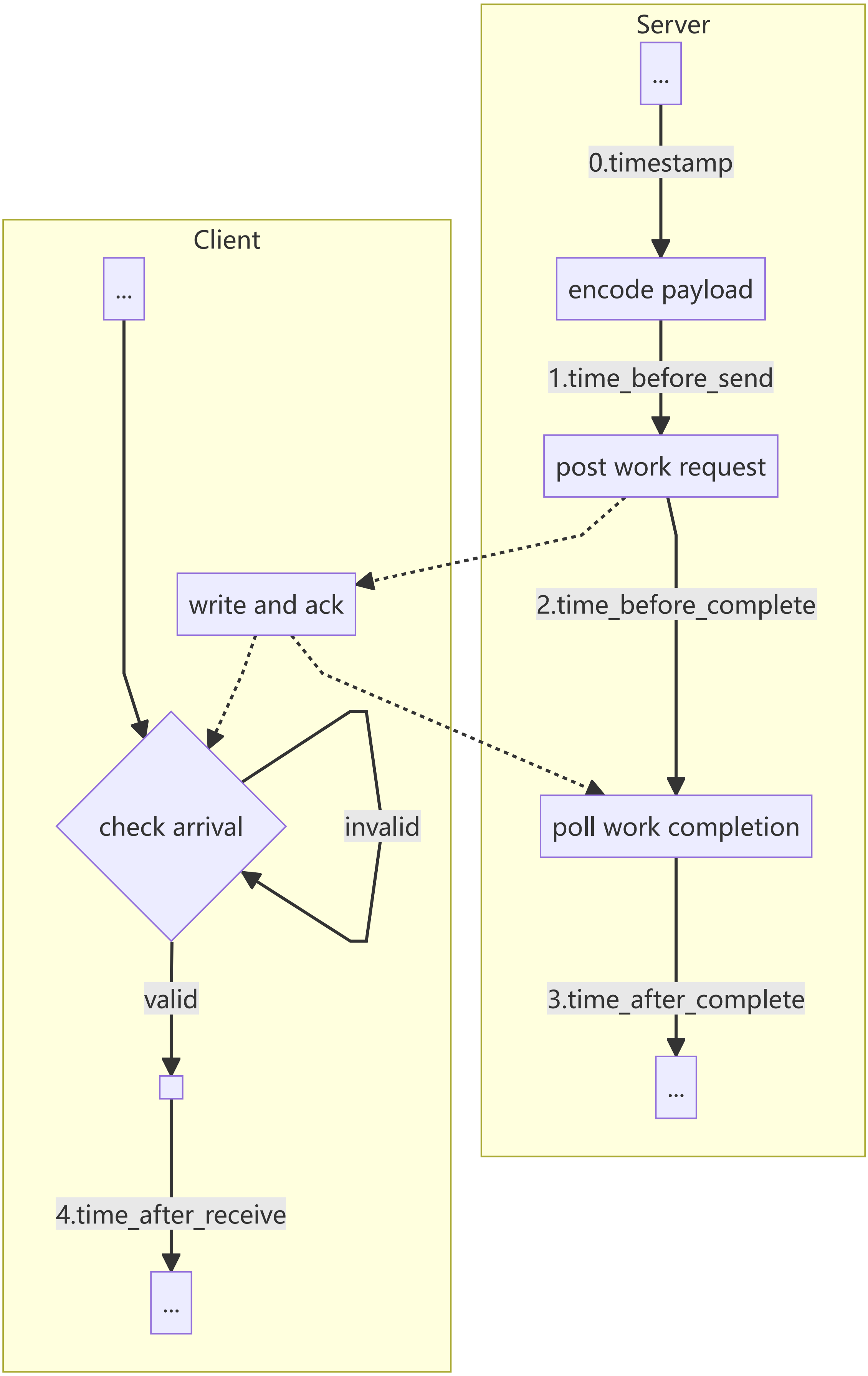
这两个估计量使用原始的统计量通过一系列滤波算法维护。

- NTP授时建立在来回链路对称的假设之上，σ为Round Trip Time。

- Client调整自身时间，使$\delta = 0$即完成时间同步。

1.2 TCS_RDMA

- 测量节点



对于每一个包的传输过程，可以测定的数据为 $T_0, T_1, \cdots, T_4$ ，我们希望找到一种方式，可以由这些数据给出与NTP协议中意义相同的统计量 σ 与 δ 。

其中 $\sigma = T_3 - T_2$ 比较方便得到，问题的关键在于如何得到 δ 。

2 理论模型

为了减小处理误差的传播，将原始时间戳数据变成每一段的值。

- $time\_encode = T_1 - T_0$
- $time\_send = T_2 - T_1$
- $time\_comp = T_3 - T_2$
- $time\_check = T_4 - T_2$

我们特别关注 $b = T_4 - T_2$ 这个统计量，这个时间段应当由3部分组成：

$$b = -\delta + l + D$$

- 其中 δ 为server领先client的时钟差
- l 为server到client的链路时延
- D 是由于client端不感知包的到达，而是周期性地check包是否到达而引入的overhead。

由于包的到达可以认为是均匀分布的，D 应当在0到 T_C 之间均匀分布，故 D 的期望应当接近 $\frac{1}{2}T_C$，其中 T_C 是client端check的周期。我们可以假设 D 的期望是由线性项和常数项组成的。

$$\bar{D} = \alpha T_C + d$$

因此有：

$$\bar{b} = -\delta + l + \alpha T_C + d$$

- 其中 l 是单向链路时延，可以用 $\frac{1}{2}\sigma = \frac{1}{2}(T_3 - T_1)$ 估计。
- $\bar{b}$ 是 $T_4 - T_2$ 的统计平均，可测得
- T_C 是可测得或可预先设定的值

因此，为了在实际场景下得到 δ，我们需要做的有以下两个事情：

- 验证上面提出的假设，即 $\bar{D}$ 与 T_C 有良好的线性关系 $\bar{D} = \alpha T_C + d$
- 验证 $\alpha \approx 0.5$，$d \approx 0$，或者验证这两个值 α 和 d 的稳定性以及找到测定这两个值的方案。

3 实验方案

3.1 实验目的

- 验证上面提出的假设，即 $\bar{D}$ 与 T_C 有良好的线性关系 $\bar{D} = \alpha T_C + d$
- 验证 $\alpha \approx 0.5$，$d \approx 0$，或者验证这两个值 α 和 d 的稳定性以及找到测定这两个值的方案。
- 验证实验设计中提出来的假设。

3.2 实验假设

实验的设计依赖于以下几个观察：

- 上文中的假设 $\bar{b} = -\delta + l + \alpha T_C + d$
- d 只与client的物理性质、负载情况有关。
- l 只与server和client之间的通路有关，我们记 l_{sc} 为当`server=s, client=c`时的链路时延。
- δ 是靠其他机制维护的量，长时间内会有较大慢变的波动项。

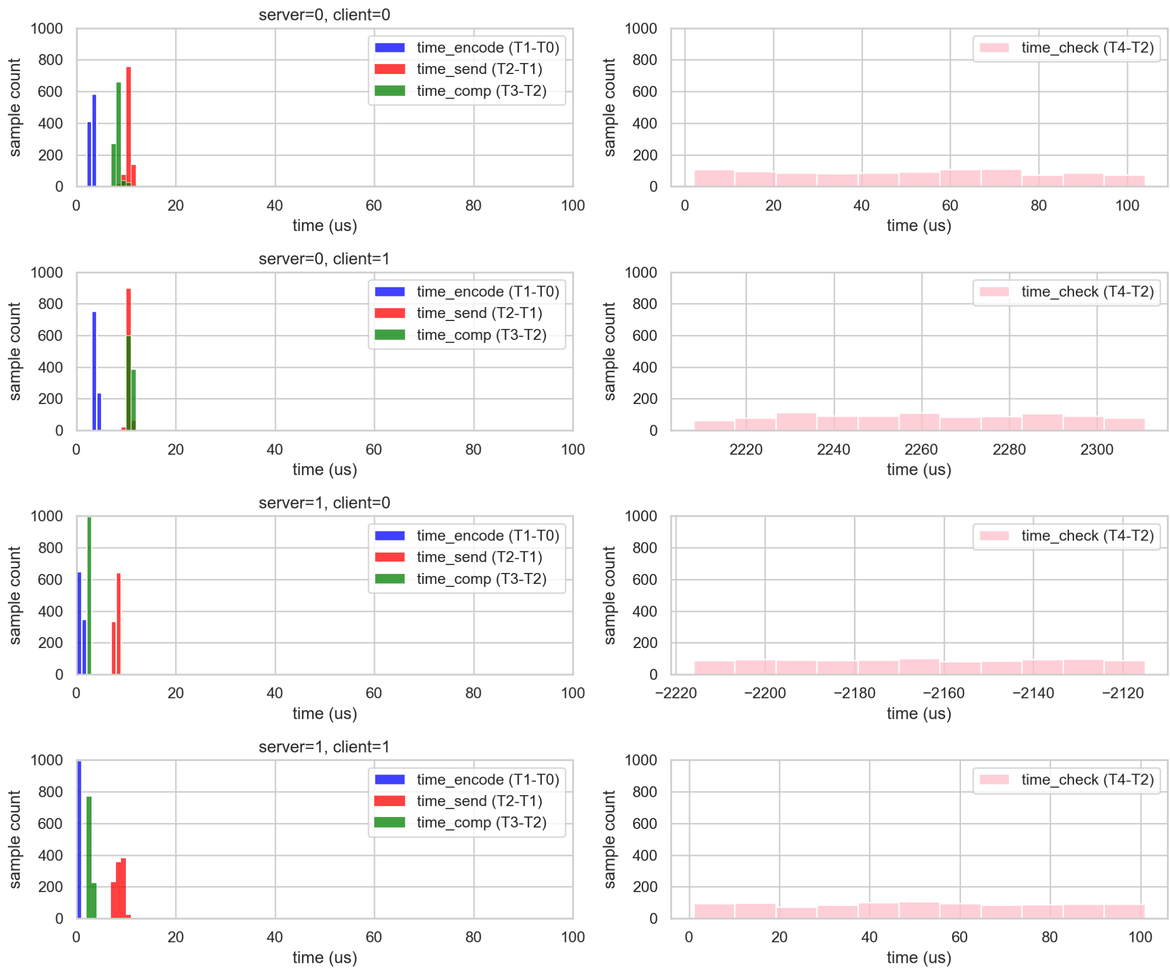
3.3 实验过程

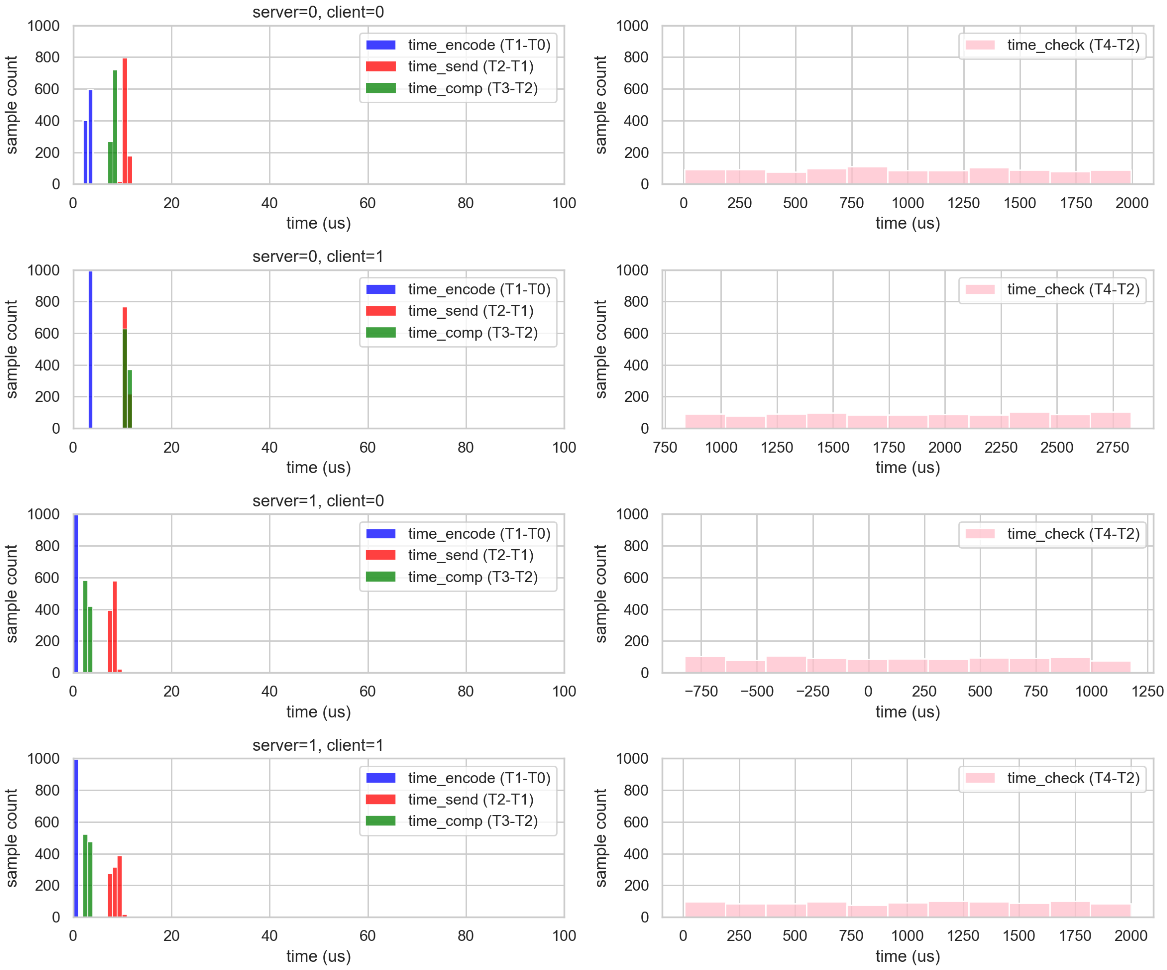
3.3.1 PDF

使用两台host0与host1之间可以搭建4条 (server,client) 的通路。对于 $T_C \in \{100, \cdots, 2000\}$ 每组固定client端的check period T_C，各组分别跑1000个点，记录下面的四个时间段的值：

- $time\_encode = T_1 - T_0$
- $time\_send = T_2 - T_1$
- $time\_comp = T_3 - T_2$
- $time\_check = T_4 - T_2$

下面是第1组 $T_C = 100\ \mu s$ 和最后一组 $T_C = 2000\ \mu s$ 的各统计量的PDF。





可以发现 $time\_check$ 基本上是在长度为 T_C 的区间均匀分布，基本上说明了 $\bar{D} = \alpha T_C + d$ 假设的背景是合理的。

下面通过这些数据的统计量验证与计算其他未知量。

3.3.2 统计量与参数关系

记实验中测得的 b_{sc} 为当`server=s, client=c`时的 $T_4 - T_2$ 的统计均值。则有以下关系：

$$
\begin{aligned}
b_{00} &= \alpha T_{C0} + l_{00} + d_0 \\
b_{01} &= \delta + \alpha T_{C1} + l_{01} + d_1 \\
b_{10} &= -\delta + \alpha T_{C0} + l_{10} + d_0 \\
b_{11} &= \alpha T_{C1} + l_{11} + d_1
\end{aligned}
$$

我们控制每一组内有 $T_{C0} = T_{C1} = T_C$，则按假设有如下关系式

$$
\begin{bmatrix}
0 & T_C & 0 & 1 & 0 & 1 & 0 \\
1 & T_C & 1 & 0 & 0 & 0 & 1 \\
-1 & T_C & 1 & 0 & 0 & 1 & 0 \\
0 & T_C & 0 & 0 & 1 & 0 & 1
\end{bmatrix}
\cdot
\begin{bmatrix}
\delta \\ \alpha \\ l_{01} \\ l_{00} \\ l_{11} \\ d_0 \\ d_1
\end{bmatrix}
=
\begin{bmatrix}
b_{00} \\ b_{01} \\ b_{10} \\ b_{11}
\end{bmatrix}
$$

化简得

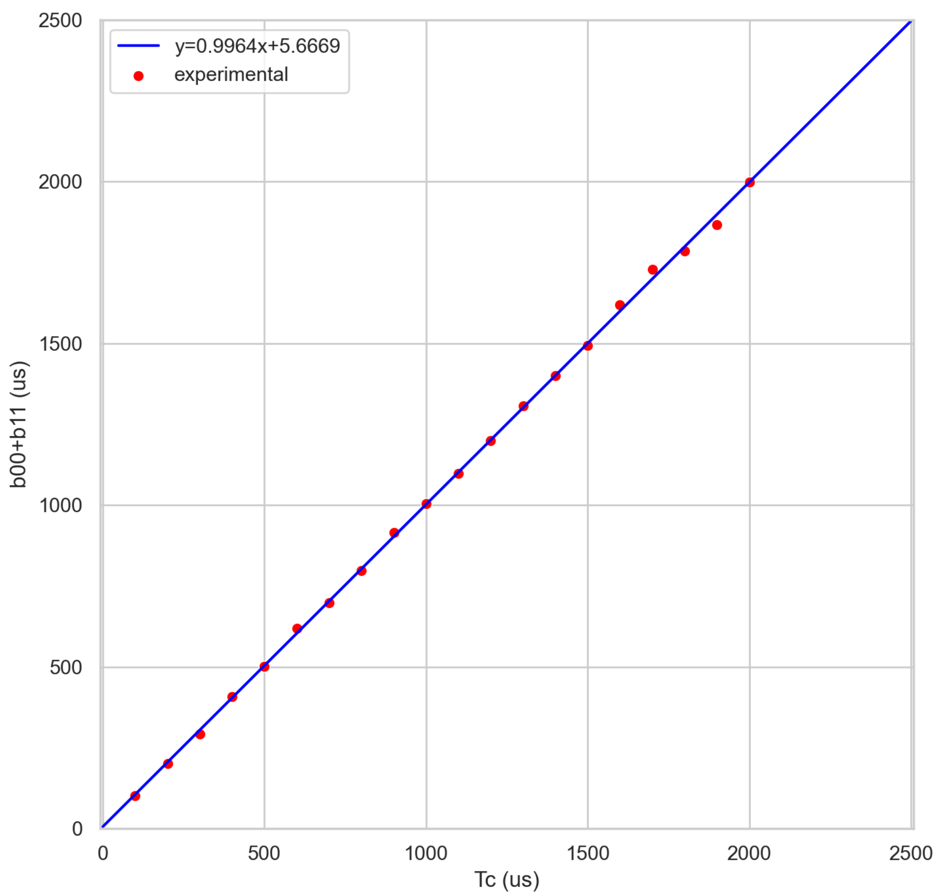
$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 2T_C & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \delta \\ \alpha \\ l_{01} \\ l_{00} \\ l_{11} \\ d_0 \\ d_1 \end{bmatrix} = \begin{bmatrix} b_{01}-b_{10} \\ b_{00}+b_{11} \\ b_{10}+b_{01}-b_{00}-b_{11} \\ b_{00}-b_{11} \end{bmatrix}$$

这是一个欠定的方程，对于不同的实验组，采用不同的 T_C 只会改变系数矩阵第2行的值，实际上可以让系数矩阵的欠定数减一。具体操作如下：

3.3.3 线性拟合

线性拟合

使用 $2T_C\alpha + (l_{00}+l_{11}+d_0+d_1) = b_{00}+b_{11}$ 进行线性拟合，一方面验证线性假设的合理性，另一方面给出斜率 2α 和截距 $l_{00}+l_{11}+d_0+d_1$ 的估计值 $\hat{\alpha}$ 和 $\hat{b}$



实验测得相关系数为 $R=0.9998$，$2\hat{\alpha}=0.9964$，$\hat{b}=5.6669$。

误差分析

可以看到 T_C 越小的点线性化程度越好，因为 $\mathbf{var}(b_{sc}) = \mathbf{var}(\delta\, or\, 0) + \frac{T_C^2}{12N}$，其中 N 是样本点数，有 $\mathbf{std}(b_{00}+b_{11}) = \frac{T_C}{\sqrt{6N}}$，可见 T_C 越大，y 轴的标准差越大，且结果可以通过增大 N 改善。

由此估计斜率的标准差为

$$\mathbf{std}(2\hat{\alpha})=\mathbf{std}\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}=\frac{1}{\sqrt{6N}}\frac{\sqrt{\sum(x_i-\bar{x})^2x_i^2}}{\sum(x_i-\bar{x})^2}$$

在这次实验中，有 $x_i\in\{100,200,\cdots,2000\}$，计算得 $\mathbf{std}(2\hat{\alpha})=\frac{0.2064}{\sqrt{N}}$，对于 $N=1000$，有 $\mathbf{std}(2\hat{\alpha})=0.0065$；对于 $N=5000$，有 $\mathbf{std}(2\hat{\alpha})=0.0029$

$$\mathbf{std}(\hat{b})=\mathbf{std}(-2\hat{\alpha}\bar{x}+\bar{y})=\mathbf{std}(\sum(\frac{1}{K}-\frac{(x_i-\bar{x})\bar{x}}{\sum(x_i-\bar{x})^2})y_i)=$$

$$\frac{1}{\sqrt{6NK}}\mathbf{std}(\sum\frac{(\sum x_i^2)-K\bar{x}x_i}{(\sum x_i^2)-K\bar{x}^2}y_i)=\frac{1}{\sqrt{6NK}}\sqrt{\sum\left(\frac{(\sum x_i^2)-K\bar{x}x_i}{(\sum x_i^2)-K\bar{x}^2}\right)^2x_i^2}$$

在此实验中，有 $K=20$，计算得 $\mathbf{std}(\hat{b})=\frac{148.84}{\sqrt{N}}$，当 $N=1000$ 时，$\mathbf{std}(\hat{b})=4.7$，当 $N=5000$ 时，$\mathbf{std}(\hat{b})=2.1$

因此上述结果较为可信。

3.3.4 检验与估计

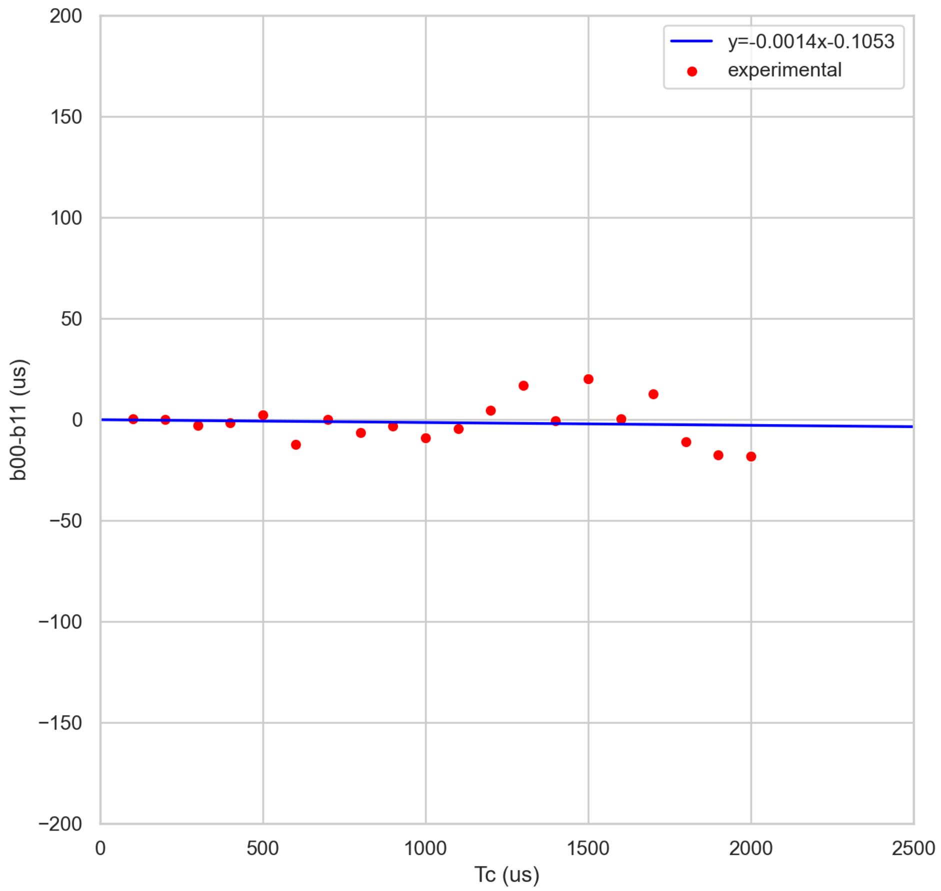
此时用估计值改写上述方程，有

$$\begin{bmatrix}2&0&0&0&-1&1\\0&0&1&1&1&1\\0&2&-1&-1&0&0\\0&0&1&-1&1&-1\end{bmatrix}\cdot\begin{bmatrix}\delta\\l_{01}\\l_{00}\\l_{11}\\d_0\\d_1\end{bmatrix}=\begin{bmatrix}b_{01}-b_{10}\\\hat{b}\\b_{10}+b_{01}-b_{00}-b_{11}\\b_{00}-b_{11}\end{bmatrix}$$

事实上，实际拟合得到的 $\hat{b}\approx l_{00}+l_{11}+d_0+d_1$ 在很小，基本上us量级。而其中的每一项都是正数，因此每一项都有一个比较小的上界。则有：

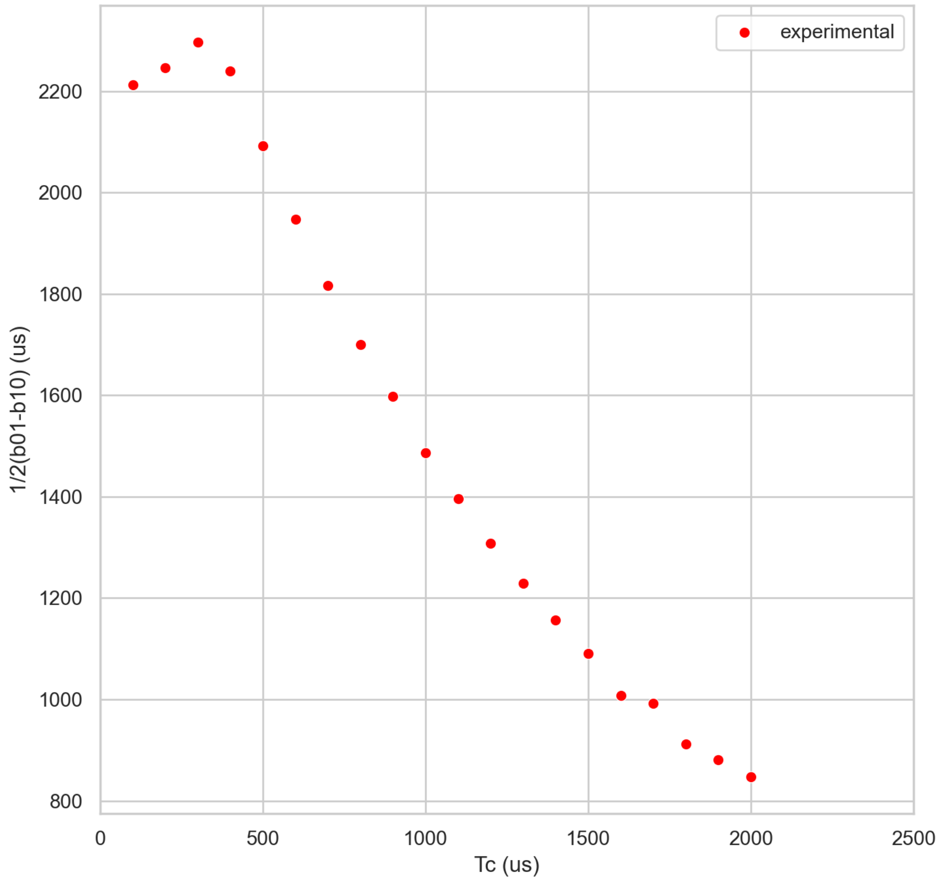
$$\begin{bmatrix}2&0\\0&0\\0&2\\0&0\end{bmatrix}\cdot\begin{bmatrix}\delta\\l_{01}\end{bmatrix}=\begin{bmatrix}b_{01}-b_{10}\\\hat{b}\\b_{10}+b_{01}-b_{00}-b_{11}\\b_{00}-b_{11}\end{bmatrix}-\begin{bmatrix}0&0&-1&1\\1&1&1&1\\-1&-1&0&0\\-1&1&-1&1\end{bmatrix}\begin{bmatrix}l_{00}\\l_{11}\\d_0\\d_1\end{bmatrix}\approx\begin{bmatrix}b_{01}-b_{10}\\\hat{b}\\b_{10}+b_{01}-b_{00}-b_{11}\\b_{00}-b_{11}\end{bmatrix}$$

其中 $\hat{b}$ 已经验证很接近0； $b_{00}-b_{11}$ 与接近0的程度可以用来验证假设。



通过剩下的方程有

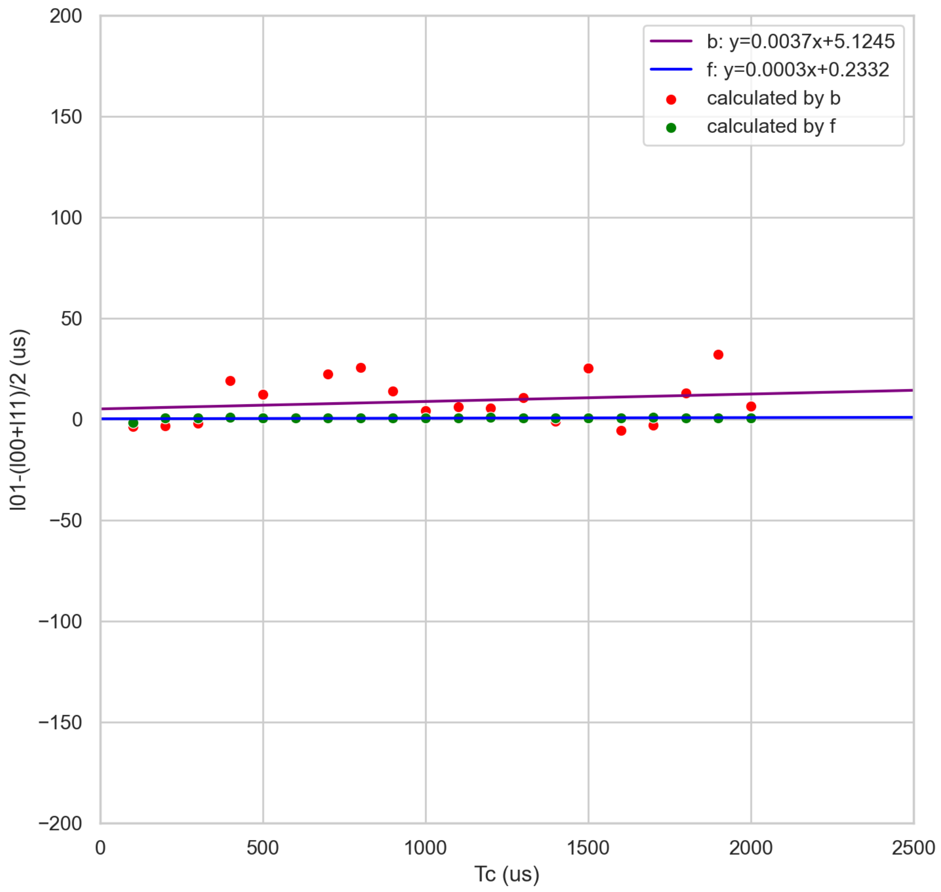
- $\delta = \frac{1}{2}(b_{01} - b_{10}) + \frac{1}{2}(d_0 - d_1)$



- $l_{01} = \frac{1}{2}(b_{10} + b_{01} - b_{00} - b_{11}) + \frac{1}{2}(l_{00} + l_{11})$

又由于 $f := (time\_send + time\_comp)$ 计算的
$l_{01} - \frac{1}{2}(l_{00} + l_{11}) = \frac{1}{4}(f_{10} + f_{01} - f_{00} - f_{11})$，可以比较两者的结果

可见由于链路很短，不能提现二者的相近之处。用 f 计算的值要稳定得多。



综上所述，主要使用的数学关系式如下：

$$
\begin{gathered}
2\alpha T_C + (l_{00} + l_{11} + d_0 + d_1) = b_{00} + b_{11} \quad \Longrightarrow \quad (\hat{\alpha}, \hat{b}) \\
\hat{\alpha} \approx 0.5, \quad \hat{b} \approx l_{00} + l_{11} + d_0 + d_1 \approx 0 \\
b_{00} - b_{11} = l_{00} - l_{11} + d_0 - d_1 \approx 0 \\
\delta = \frac{1}{2}(b_{01} - b_{10}) + \frac{1}{2}(d_0 - d_1) \approx \frac{1}{2}(b_{01} - b_{10}) \\
l_{01} - \frac{1}{2}(l_{00} + l_{11}) = \frac{1}{2}(b_{10} + b_{01} - b_{00} - b_{11}) \\
l_{01} - \frac{1}{2}(l_{00} + l_{11}) = \frac{1}{4}(f_{10} + f_{01} - f_{00} - f_{11})
\end{gathered}
$$

4 重复实验

4.1 实验设定

做了4次实验，每次实验的设定有所不同，以上部分的结果都是第一次实验的结论。

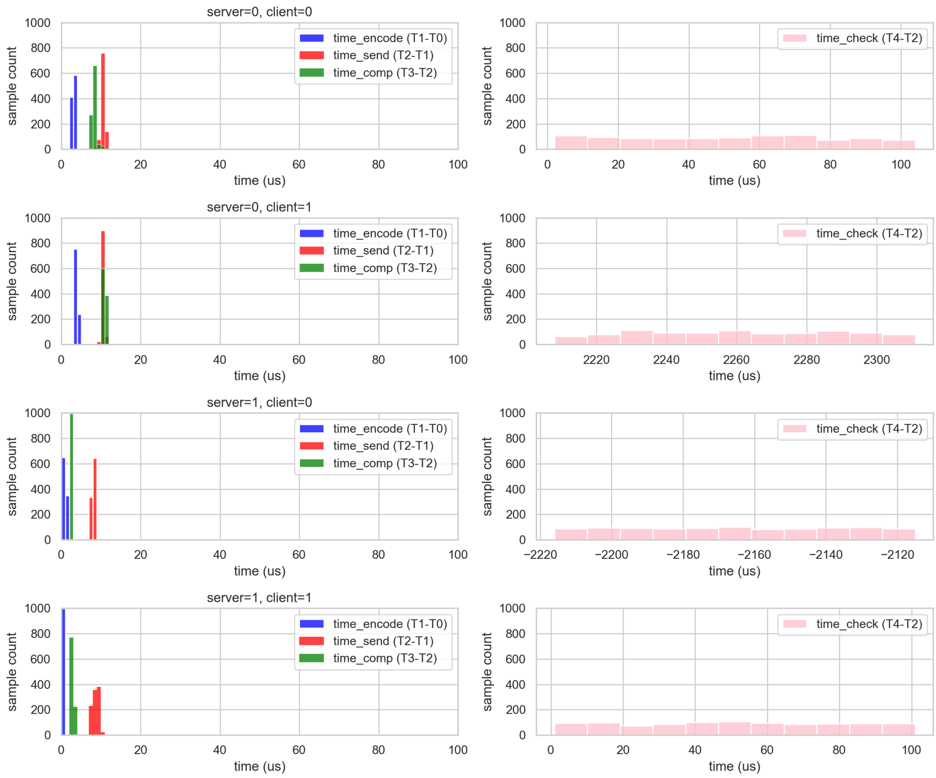
	实验一	实验二	实验三	实验四
T_C取值范围	100:100:2000	100:100:2000	100:100:2000	100:100:2000
host0	192.168.1.70	192.168.1.70	192.168.1.70	192.168.1.70

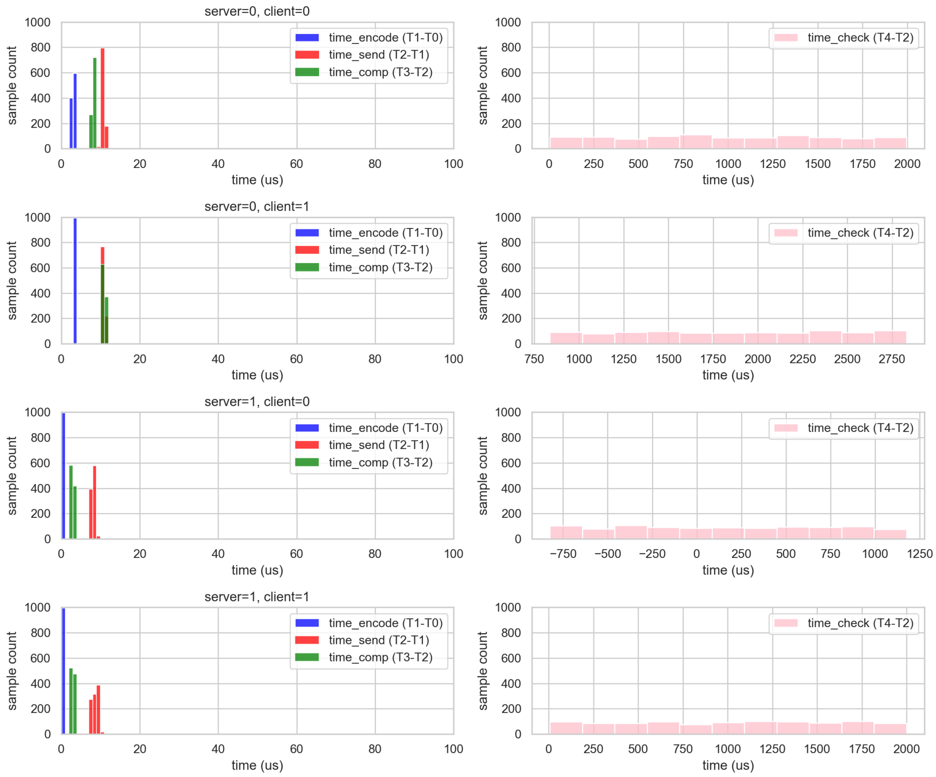
	实验一	**实验二**	**实验三**	**实验四**
host1	192.168.1.71	192.168.1.71	192.168.1.71	192.168.1.71
每个实验组发送的点数	1000	1000	1000	5000
时间跨度	~ 640 s	~ 640 s	~ 35 s	~ 55 s
说明	最初的数据	实验一做完几个小时后跑的数据	将每个T_C对应的实验组改并行启动	扩大样本点数查看收敛性能变化

以下是一些对等数据的比较

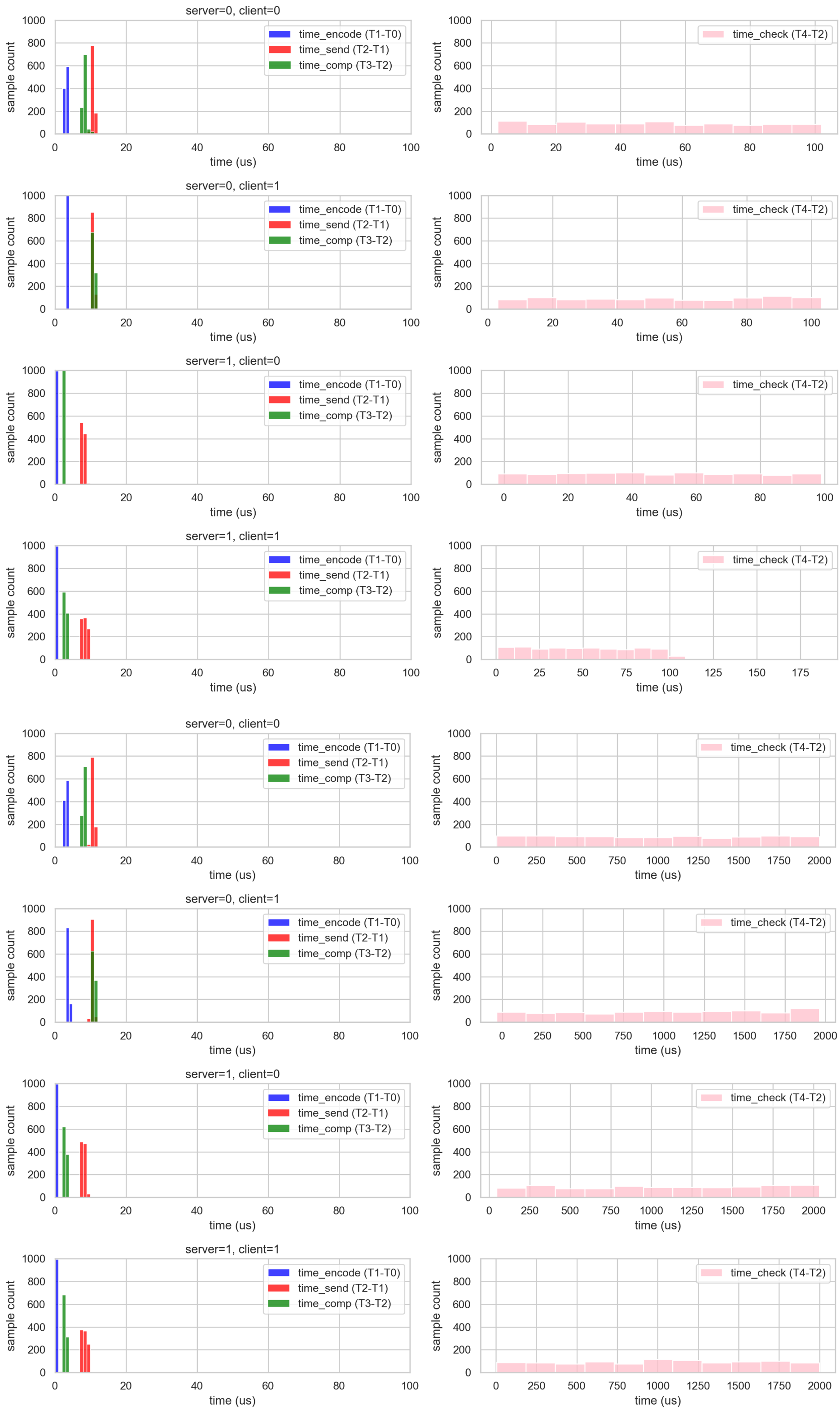
4.2 $T_C = 100, 2000$ 的PDF特征

- **实验一**

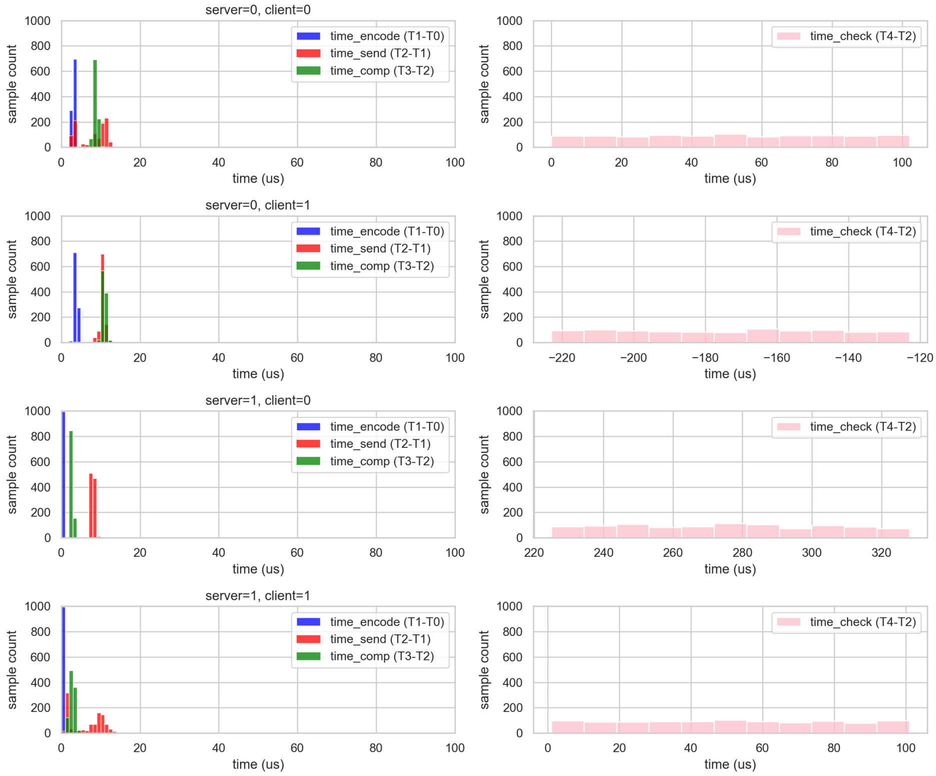

server=0, client=0
server=0, client=1
server=1, client=0
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
sample count
time (us)

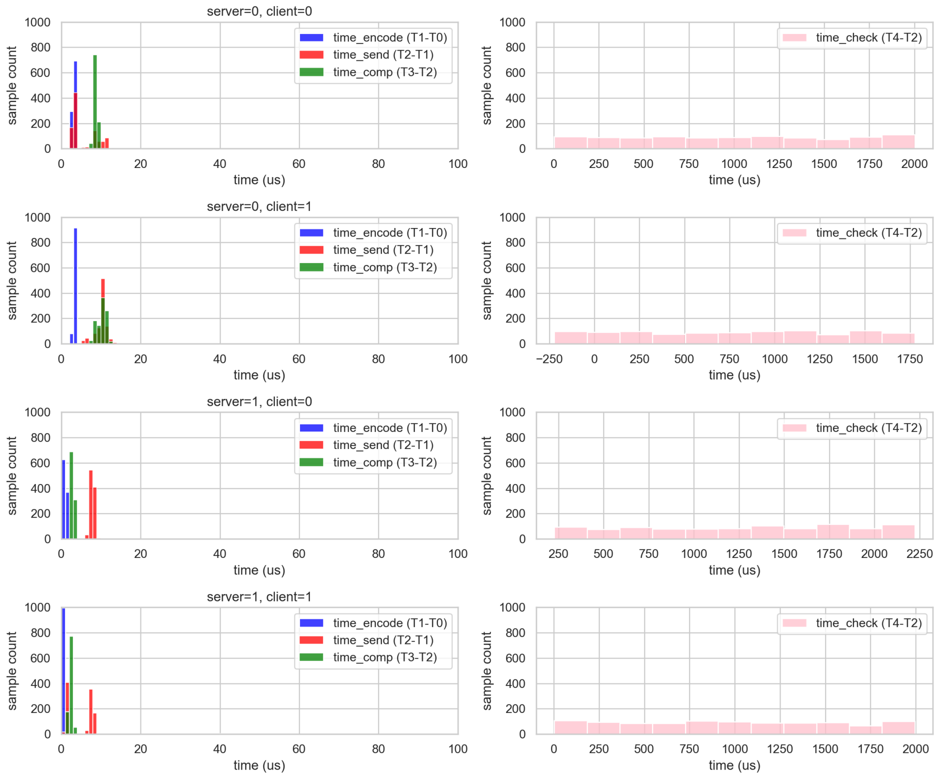

server=0, client=0
server=0, client=1
server=1, client=0
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
sample count
time (us)

- **实验二**

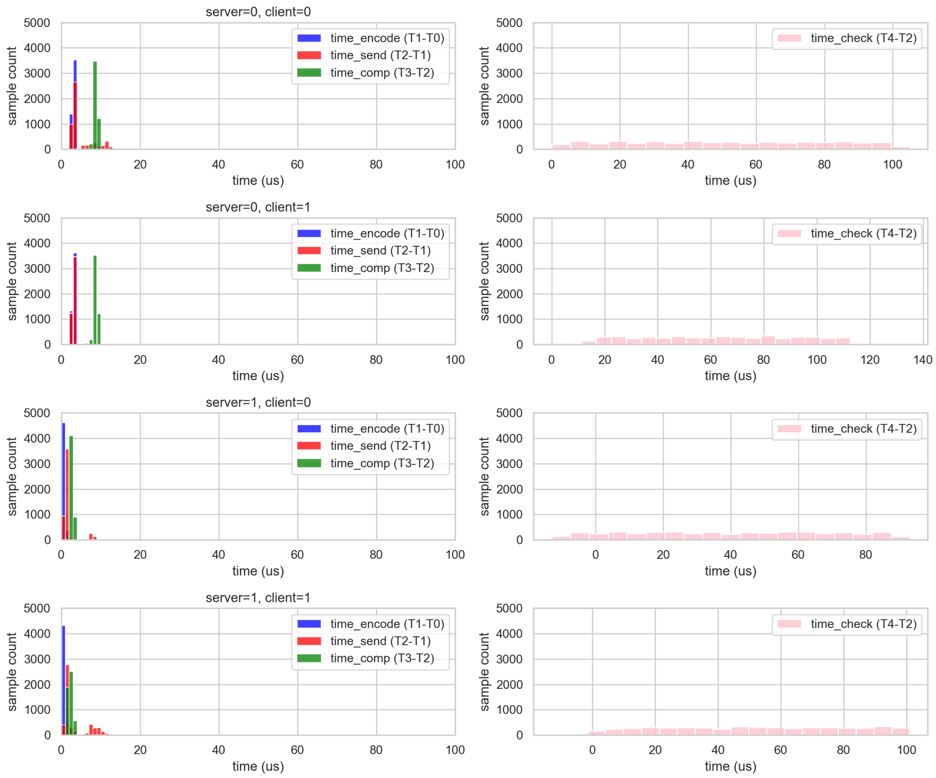

server=0, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=0, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=0, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=0, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
sample count
time (us)

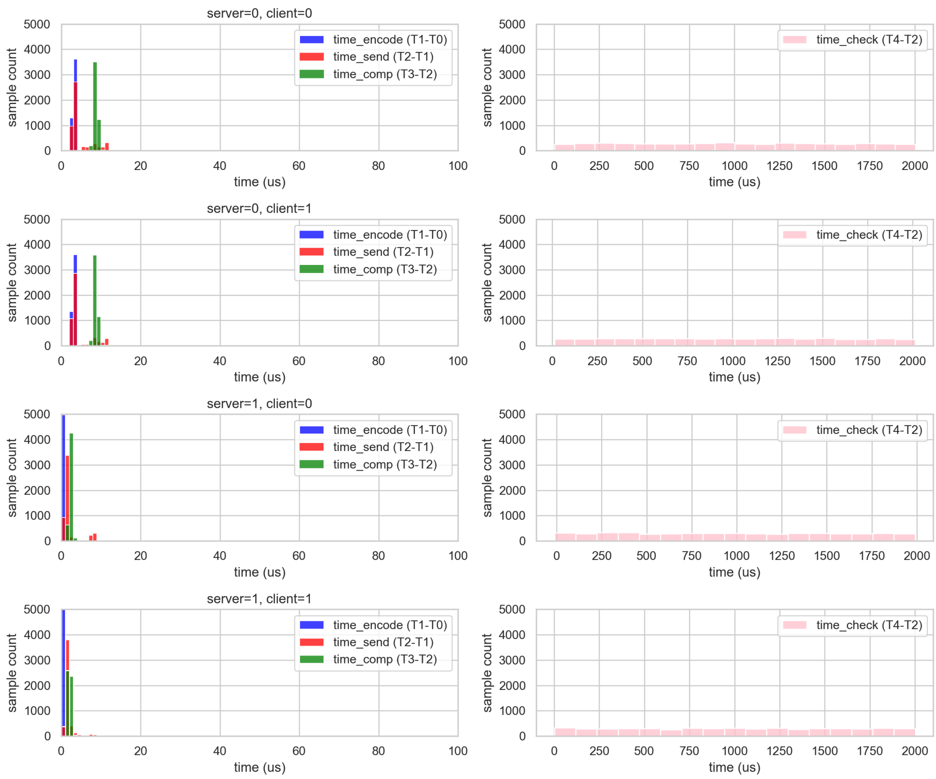
- **实验三**


server=0, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=0, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
sample count
time (us)


server=0, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=0, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=0
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
server=1, client=1
time_encode (T1-T0)
time_send (T2-T1)
time_comp (T3-T2)
time_check (T4-T2)
sample count
time (us)

- **实验四**

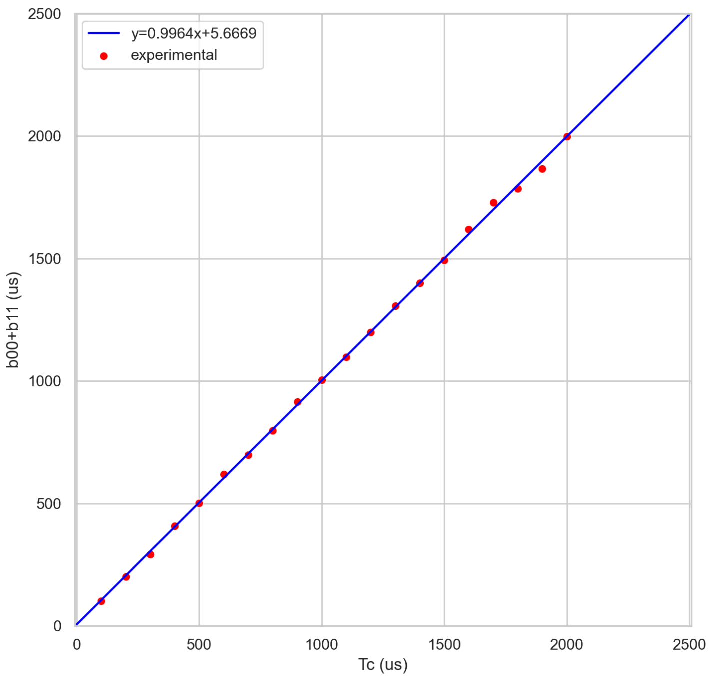


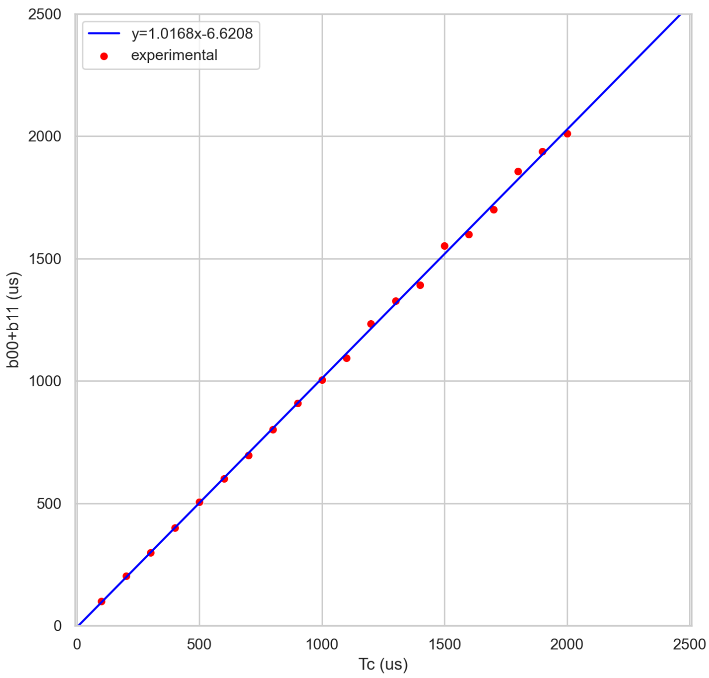


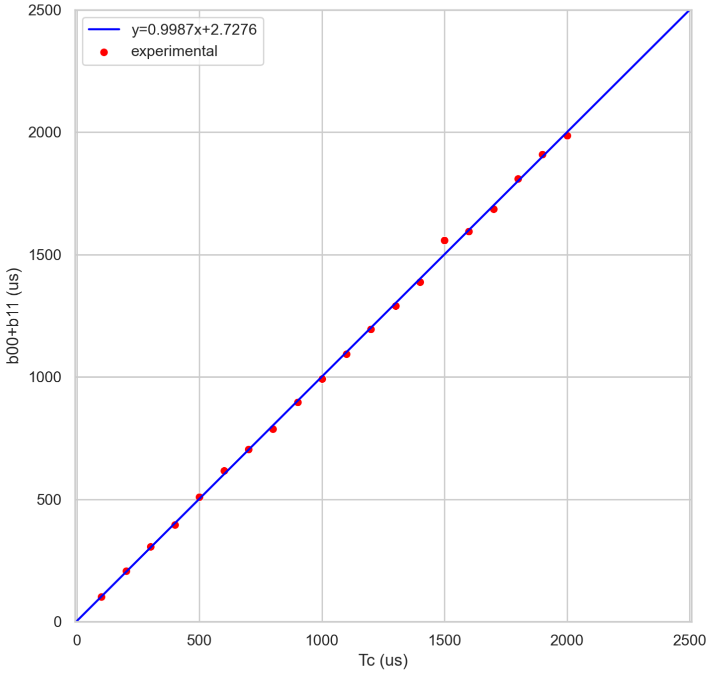
4.3 α 的曲线拟合

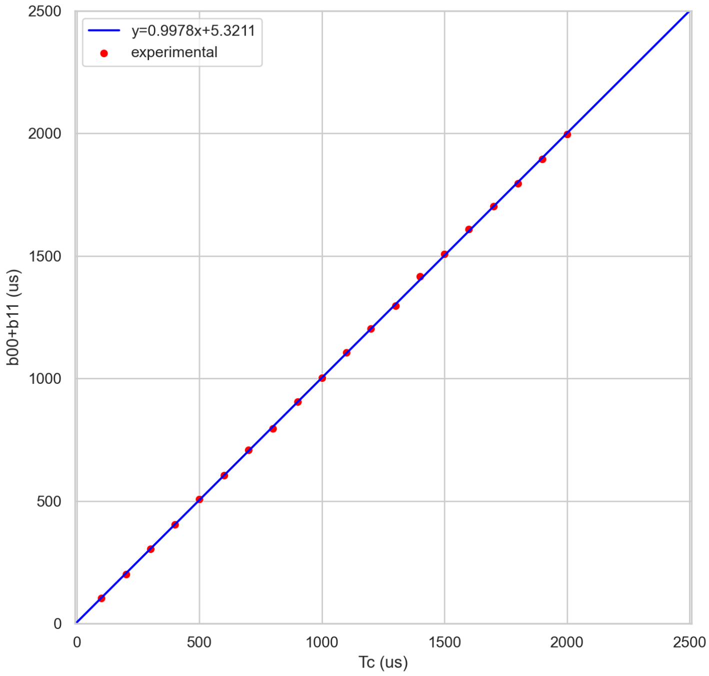
以下图中的子图顺序如下表所示

实验一	实验二
实验三	实验四


y=0.9964x+5.6669
experimental
b00+b11 (us)
Tc (us)
0
500
1000
1500
2000
2500

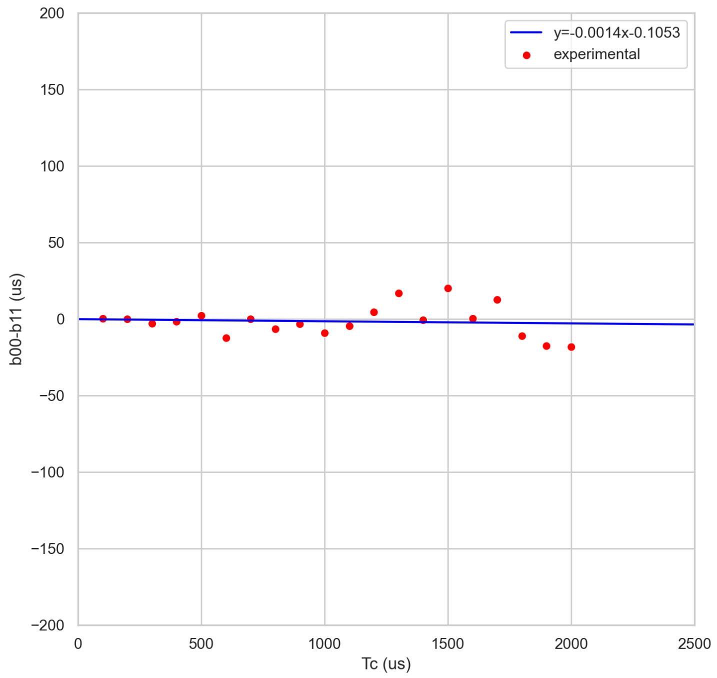

y=1.0168x-6.6208
experimental
b00+b11 (us)
Tc (us)
0
500
1000
1500
2000
2500

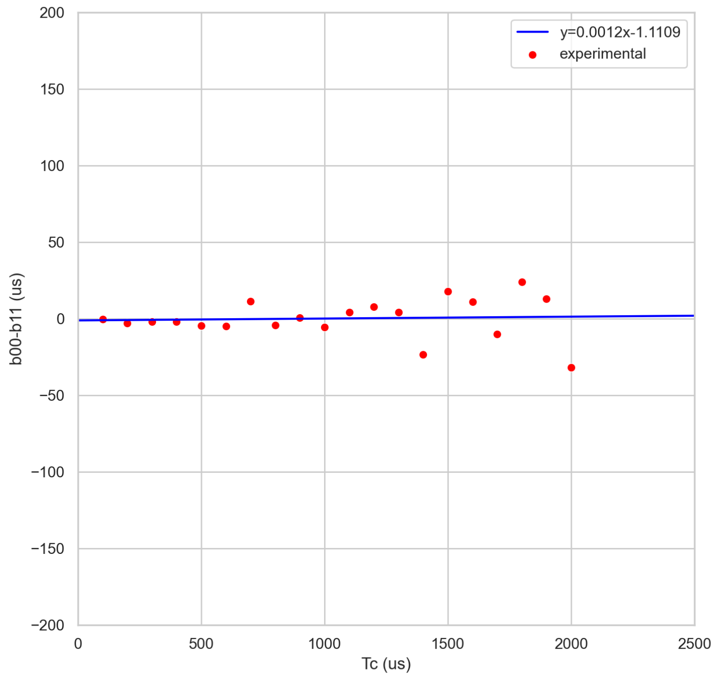


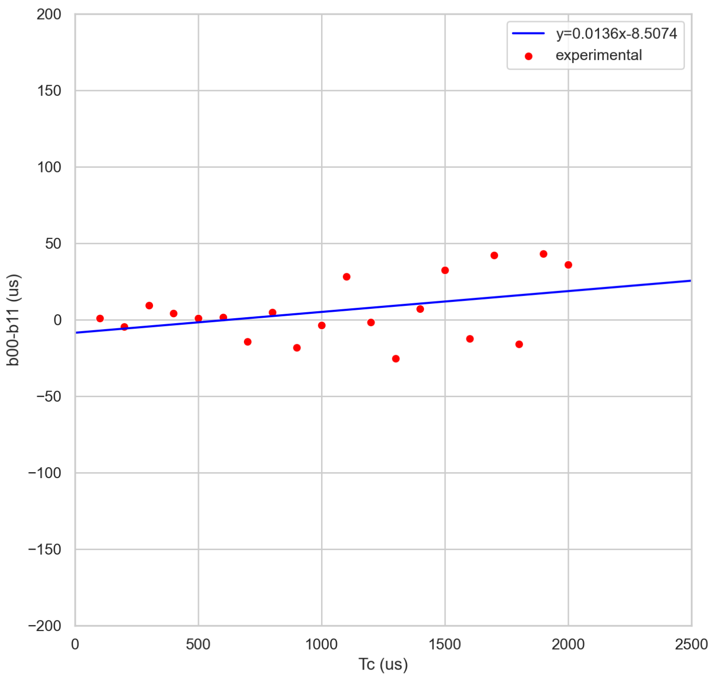


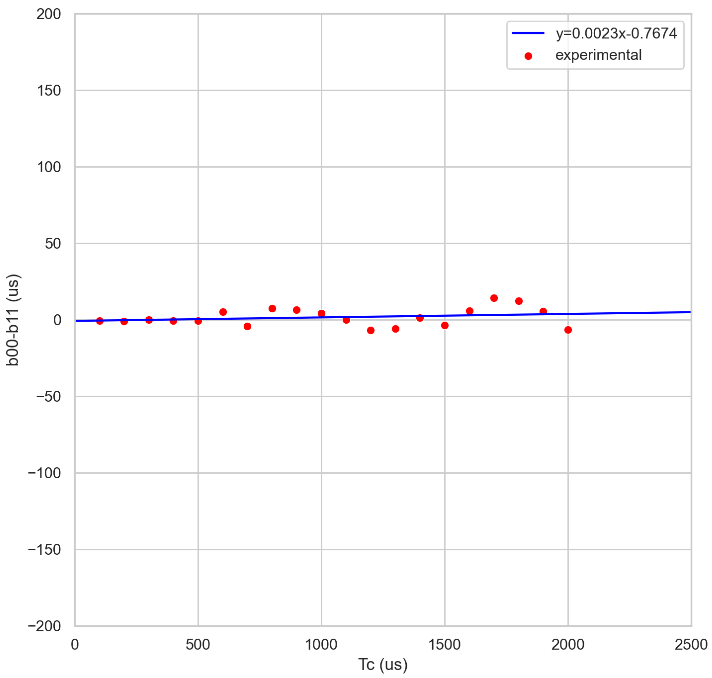
	实验一	实验二	实验三	实验四
相关系数	0.99976	0.99962	0.99962	0.99996
斜率	0.9964	1.0168	0.9987	0.9978
截距	5.6669	-6.6208	2.7276	5.3211

4.4 $b_{00} - b_{11}$的近0检验

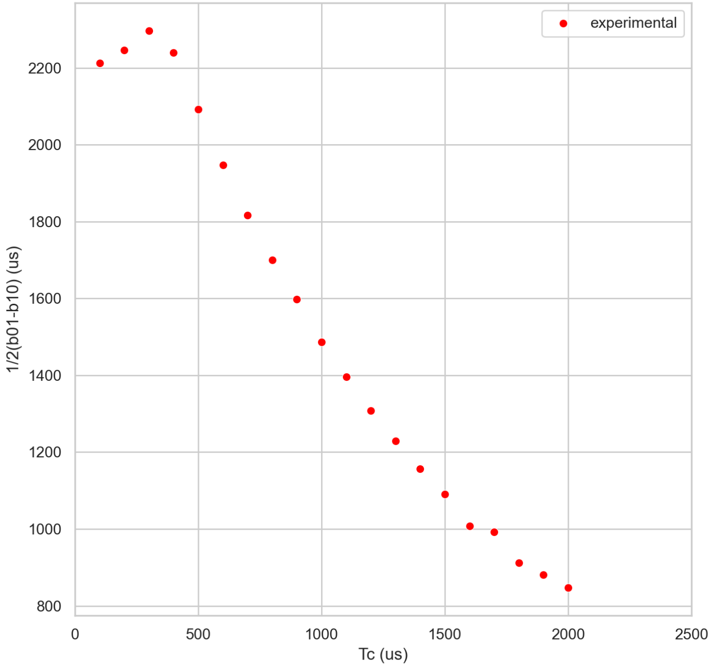


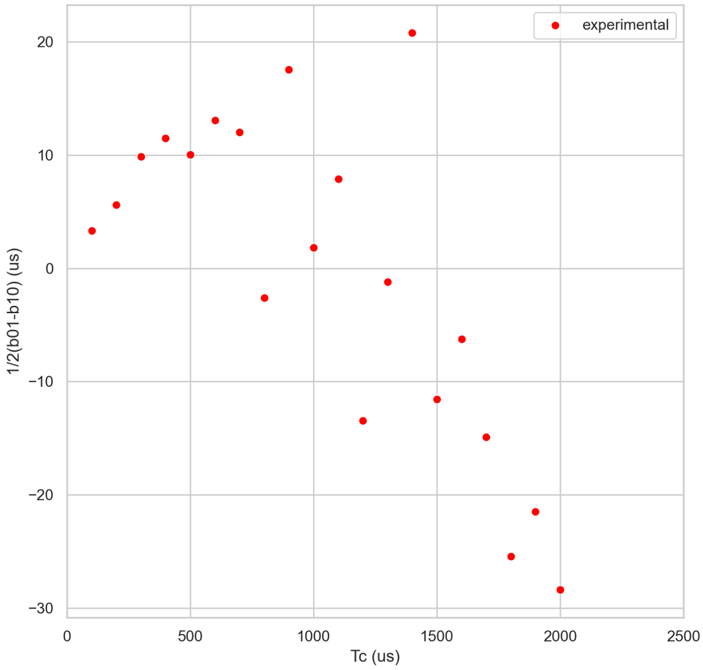


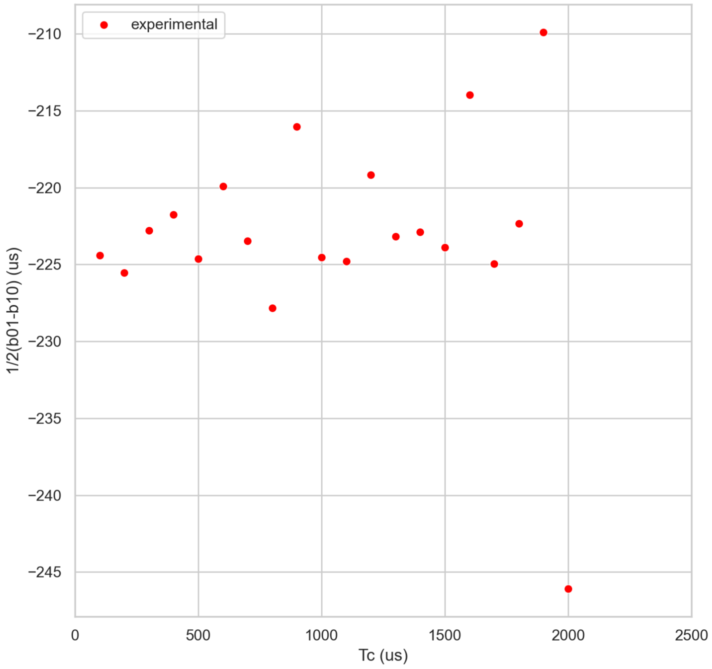

y=0.0136x-8.5074
experimental
b00-b11 (us)
Tc (us)

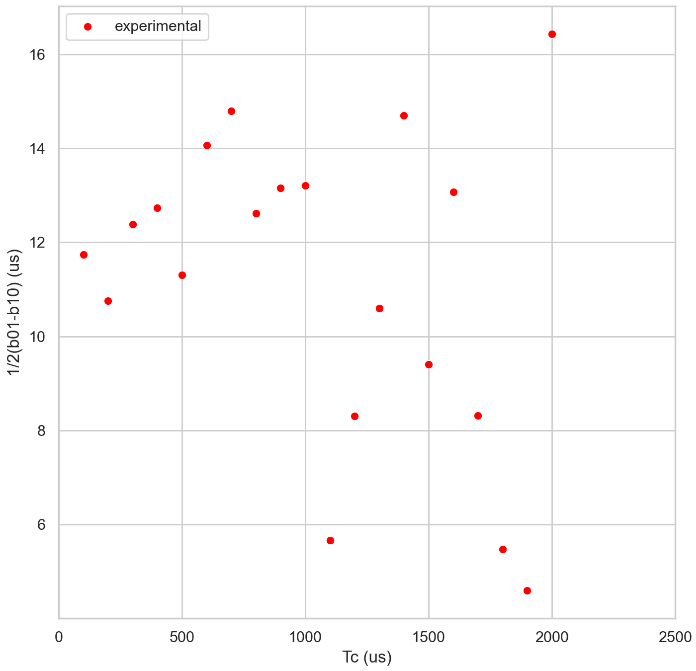

y=0.0023x-0.7674
experimental
b00-b11 (us)
Tc (us)

4.5 δ 的估计

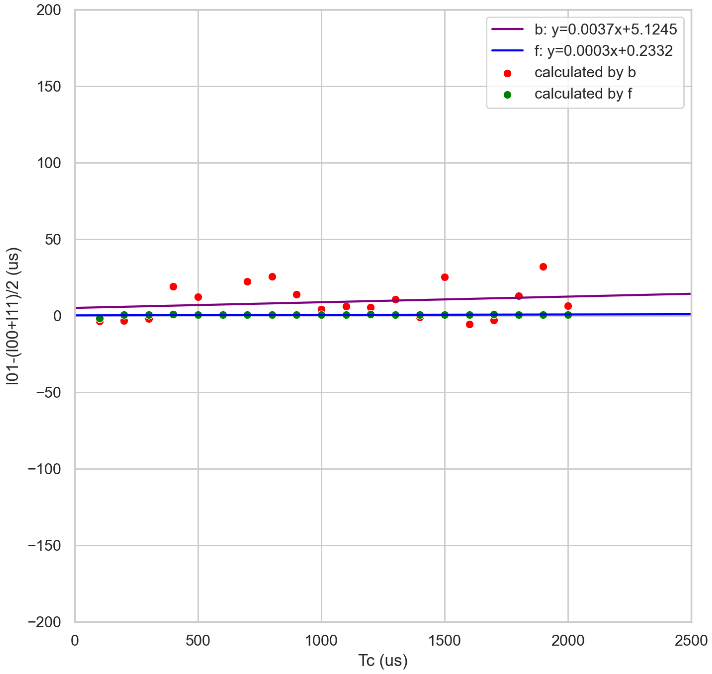

experimental
2200
2000
1800
1600
1400
1200
1000
800
1/2(b01-b10) (us)
0
500
1000
1500
2000
2500
Tc (us)

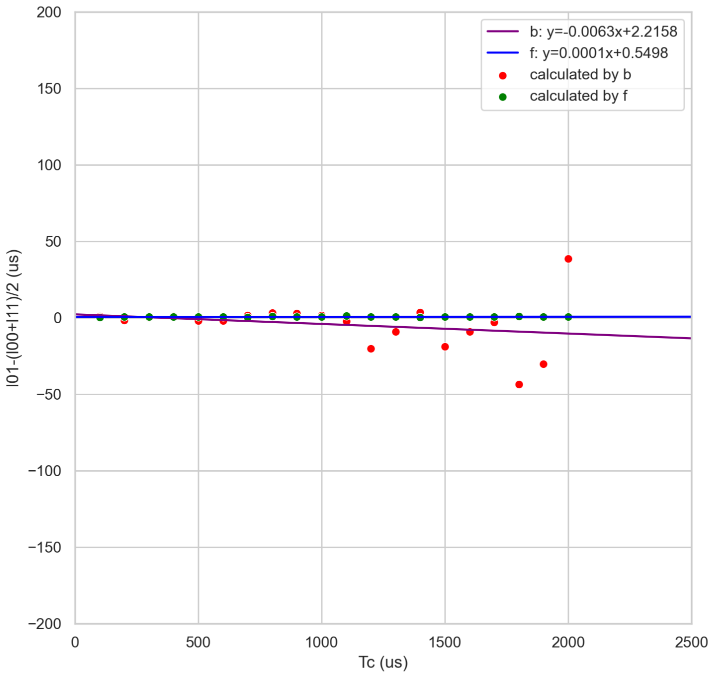

experimental
20
10
0
−10
−20
−30
1/2(b01-b10) (us)
0
500
1000
1500
2000
2500
Tc (us)

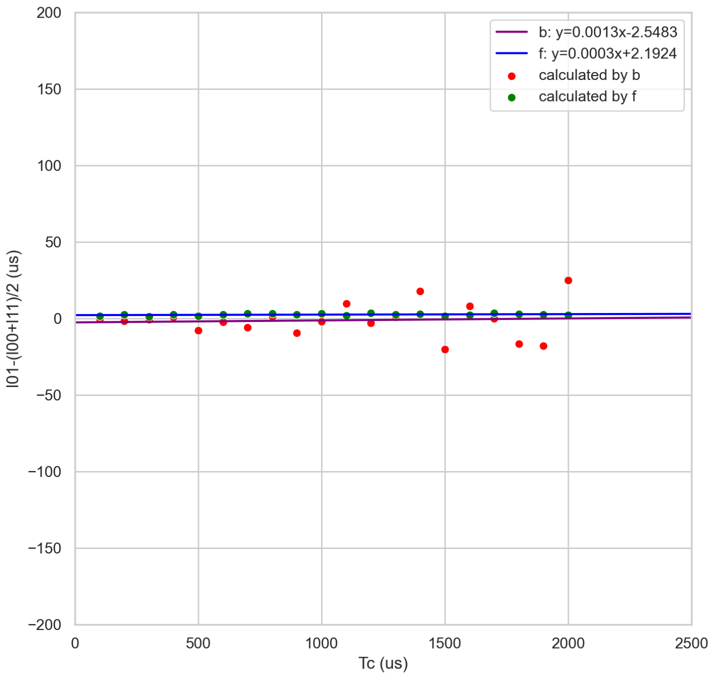

experimental
−210
−215
−220
−225
−230
−235
−240
−245
1/2(b01-b10) (us)
0
500
1000
1500
2000
2500
Tc (us)

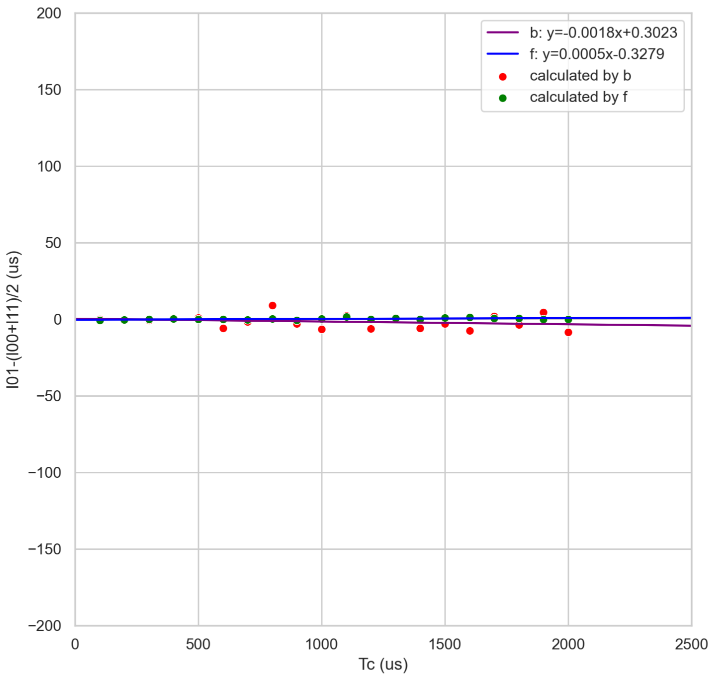

experimental
16
14
12
10
8
6
1/2(b01-b10) (us)
0
500
1000
1500
2000
2500
Tc (us)

4.6 $l_{01}-\frac{1}{2}(l_{00}+l_{11})$ 的估计与检验









5 实验结论

- 验证了假设模型 $\bar{b} = -\delta + l + \alpha T_C + d \approx -\delta + l + \frac{1}{2} T_C$
- 来回方案：可以通过server与client之间分别作为主动方传送数据测定延时，用 $\delta \approx b_{01} - b_{10}$ 估计 δ，可以得到单点精度被 T_C 控制，均值精度在 μs 级别的 δ
- 单边方案：可以通过 $\sigma = T_3 - T_2$ 估计得到 $l_{01} \approx \frac{1}{2}\sigma$，可以直接通过测定 T_C, T_2, T_4，用 $\delta \approx l_{01} + \frac{1}{2} T_C - (T_4 - T_2)$ 可以得到单点精度被 $\frac{1}{2} T_C$ 控制，均值精度在 μs 级别的 δ，此外精度还会受到 l_{01} 估算精度的影响，l_{01} 的精度需要后面的实验确定，目前来看单点误差最多是 $10\mu s$ 量级。